

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

CARLO CORRÊA SOLCI

ESTUDO EMPÍRICO DE MÉTODOS DE
ESTIMAÇÃO ROBUSTA DE MODELOS PAR E
APLICAÇÃO A DADOS DE POLUIÇÃO DO AR

VITÓRIA
2017

CARLO CORRÊA SOLCI

**ESTUDO EMPÍRICO DE MÉTODOS DE ESTIMAÇÃO ROBUSTA DE
MODELOS PAR E APLICAÇÃO A DADOS DE POLUIÇÃO DO AR**

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do título de Mestre em Engenharia Ambiental, na área de concentração Poluição do Ar.

Orientador: Prof. Dr. Valdério Anselmo Reisen.

Coorientador: Prof. Dr. Alessandro José Queiroz Sarnaglia.

VITÓRIA

2017

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)
Bibliotecário: Fábio Massanti Medina – CRB-6 ES 000629/O

S684e Solci, Carlo Corrêa, 1990-
Estudo Empírico de Métodos de Estimação Robusta de
Modelos PAR e Aplicação a Dados de Poluição do Ar / Carlo
Corrêa Solci. – 2017.
57 f. : il.

Orientador: Valdério Anselmo Reisen
Coorientador: Alessandro José Queiroz Sarnaglia
Dissertação (Mestrado em Engenharia Ambiental) –
Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Ar – Poluição. 2. Análise de séries temporais. 3. Ciclos. I.
Reisen, Valdério Anselmo. II. Sarnaglia, Alessandro José
Queiroz. III. Universidade Federal do Espírito Santo. Centro
Tecnológico. IV. Título.

CDU: 628



UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

“Estudo Empírico de Métodos de Estimação Robusta de Modelos PAR e Aplicação a Dados de Poluição do Ar”

Carlo Correa Solci

Banca Examinadora:

Prof. Dr. Valdério Anselmo Reisen
Orientador - DEA/CT/UFES

Prof. Dr. Alessandro José Queiroz Sarnaglia
Coorientador – CCE/UFES

Prof. Dr. Neyval Costa Reis Júnior
Examinador Interno - DEA/CT/UFES

Prof. Dr. Marinho Gomes de Andrade Filho
Examinador Externo – ICMC/USP

Prof. Dr. Fábio Júlio da Silva Valentim
Examinador Externo – CCE/UFES

Prof. Dr. Diogo Costa Buarque Coordenador do PPGEA:
Programa de Pós-Graduação em Engenharia Ambiental
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Vitória-ES, 09 de novembro 2017.

Agradecimentos

- ao meu orientador, Valdério, pela dedicação, orientação precisa, conhecimentos compartilhados, envolvimento na pesquisa, confiança e amizade;
- ao meu coorientador, Alessandro, pelo apoio em todos os momentos desde trabalho, pelos conhecimentos transmitidos, confiança e amizade;
- à minha tia Jane Therezinha: sua ausência física se transformou em presença espiritual na minha vida (*In Memoriam*);
- à toda a minha família que sempre me apoiou em todos os momentos da minha vida;
- à Maria da Penha dos Santos pelo apoio logístico e amizade ao longo de toda a minha vida;
- à professora Jane Santos, pelas orientações e pelos conhecimentos transmitidos;
- aos professores da banca examinadora Marinho, Neyval e Fabio pelas sugestões e críticas a este trabalho;
- aos demais professores, coordenadores e funcionários do PPGEA, pelo apoio;
- aos colegas Paulo, Carolina e Fátima pelo apoio, e em especial ao Wanderson, Edson e Higor pela incessante ajuda;
- aos demais alunos do NUMES e do PPGEA.
- aos demais professores do PPGEA e à todos os funcionários do programa.

“All models are wrong, but some are useful”
(George Box)

Lista de Figuras

Figura 1 – Série temporal em estudo após transformação logarítmica com box-plot e gráficos de dispersão em defasagem unitária para diferentes dias da semana.	15
Figura 2 – Localização espacial das estações da RAMQAr.	31
Figure 3 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 0$, $n = 400$ and normal errors.	45
Figure 4 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 7$, $n = 400$ and normal errors.	46
Figure 5 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 0$, $n = 400$ and asymmetric errors.	47
Figure 6 – Evolution of the $\log(\text{PM}_{10})$ time series.	47
Figure 7 – Periodogram of the $\log(\text{PM}_{10})$ time series.	48
Figure 8 – Daily box-plots of the $\log(\text{PM}_{10})$ time series.	48
Figure 9 – ACF of the residuals of the YWE fit.	51
Figure 10 – ACF of the residuals of the RYWE fit.	51
Figure 11 – ACF of the residuals of the RLSE fit.	51

Lista de Tabelas

Tabela 1 – Valores comuns de λ e suas transformações associadas.	28
Tabela 2 – Poluentes e parâmetros meteorológicos em cada estação da RAMQAr. .	31
Table 3 – Parameters of PAR(1) models used in the simulation.	41
Table 4 – Bias and RMSE for Model 1 and outliers with probability $\xi = 0.01$. . .	42
Table 5 – Bias and RMSE for Model 2 and outliers with probability $\xi = 0.01$. . .	43
Table 6 – Estimated coefficients of the linear model.	49
Table 7 – Selected AR orders using the BIC.	49
Table 8 – Estimates of the AR coefficients for YWE, RYWE and RLSE.	50
Table 9 – Fitting performance of the estimated models.	50

Resumo

Os estudos de poluição atmosférica geralmente envolvem medições e análises de dados de concentrações de poluentes, como é o caso do MP_{10} (material particulado de diâmetro inferior a $10\text{ }\mu\text{m}$), do SO_2 (dióxido de enxofre) e de outros poluentes. Esses dados normalmente possuem características estatísticas importantes como autocorrelação, sazonalidade, periodicidade e a presença de picos na série que apesar de não serem observações atípicas (outliers) pela alta frequência com a qual ocorrem, podem ser modelados como tais pelo efeito que têm na série. Todas essas características exigem atenção especial durante a análise dos dados. Com esse motivação, esse estudo comparou o desempenho dos estimadores robustos para modelos periódicos autorregressivos (PAR) propostos por Sarnaglia, Reisen & Lévy-Leduc (2010) (Yule-Walker Robusto) e Shao (2008) (Mínimos Quadrados Robusto), por meio de um estudo de Monte Carlo em diferentes cenários, incluindo: contaminação por observações atípicas aditivas e desvios da normalidade. Para efeito de comparação, também foi considerada a metodologia clássica de Yule-Walker que pode ser vista em McLeod (1994), por exemplo. O interesse prático em poluição do ar é avaliar se ambas as metodologias robustas captam melhor a estrutura de correlação da série do que a metodologia clássica. Isso pode ser verificado, por exemplo, pela ordem dos modelos autorregressivos obtidas por cada procedimento de estimação. Esses três métodos foram aplicados para o ajuste do modelo PAR à dados de MP_{10} da estação da Enseada do Suá da rede de monitoramento da qualidade do ar da Grande Vitória-ES.

Palavras-chave: poluição do ar, análise de séries temporais, sazonalidade, periodicidade, observações atípicas, robustez e MP_{10} .

Abstract

Studies about air pollution typically involve measurements and analysis of pollutants, such as PM_{10} (particulate matter with diameter lower than $10\text{ }\mu\text{m}$), SO_2 (sulfur dioxide) and others. These data typically have important features like serial correlation, seasonality, periodicity and the presence of peaks that despite not being atypical observations (outliers) because of their high frequency of occurrence, can be modeled as such owing to the effect that they have on the series. All these features demand special attention during data analysis. With this motivation, this study compared the performance of the robust estimator of periodic autoregressive (PAR) models proposed by Sarnaglia, Reisen & Lévy-Leduc (2010) (Robust Yule-Walker) and Shao (2008) (Robust Least Squares), through a Monte Carlo study under different scenarios, including: additive outliers contamination and departures from normality. For comparison purposes, we have also considered the classical periodic Yule-Walker methodology which can be seen in McLeod (1994), for example. The practical purpose in air pollution is to evaluate if both robust methodologies are more efficient in capturing the correlation structure than the classic one. This can be checked, for instance, by the order of the autoregressive models obtained by each estimation procedure. These three methods were applied to adjust a PAR model to MP_{10} data of Ensada do Suá station of the air quality monitoring network of the Greater Vitória Region - ES.

Keywords: air pollution, time series analysis, seasonality, periodicity, outliers, robustness and PM_{10} .

Sumário

1	INTRODUÇÃO	13
2	OBJETIVOS	17
2.1	Objetivo Geral	17
2.2	Objetivos Específicos	17
3	REVISÃO BIBLIOGRÁFICA	18
4	CONCEITOS FUNDAMENTAIS	21
4.1	Processo Estocástico	21
4.2	Processos Estocásticos Estacionários	21
4.3	Funções de Autocovariância e de Autocorrelação	22
4.4	Função de Autocorrelação Parcial	23
4.5	Processos Ruído Branco	23
4.6	Modelos Estacionários de Séries Temporais	24
4.6.1	Representações de Médias Móveis e Autorregressiva de Modelos de Séries Temporais	24
4.6.2	Modelo Autorregressivo	24
4.6.3	Modelo de Médias Móveis	25
4.6.4	Modelo Autorregressivo de Médias Móveis	25
4.7	Sazonalidade	26
4.8	Modelos Não Estacionários de Séries Temporais	26
4.8.1	Modelos de Tendência Determinísticos	27
4.8.2	Modelos Autoregressivos Integrados de Médias Móveis	27
4.9	Transformações Estabilizadoras da Variância	28
4.10	Correlação Periódica	29
5	MATERIAIS E MÉTODOS	30
5.1	Região de Estudo	30
5.2	Rede Automática de Monitoramento da Qualidade do Ar	30
5.3	Dados	32
5.4	Software Estatístico	32
6	EMPIRICAL STUDY OF ROBUST ESTIMATION METHODS FOR PAR MODELS WITH APPLICATION TO PM₁₀ DATA	33
6.1	Introduction	33

6.2	The PAR Model and its Estimation Methods	36
6.2.1	PAR Model	36
6.2.2	Estimation Methods	36
6.2.2.1	Yule-Walker Estimator	37
6.2.2.2	Robust Yule-Walker Estimator	38
6.2.2.3	Robust Least Squares Estimator	39
6.3	Simulation	41
6.3.1	Specification of the Monte Carlo Experiment	41
6.3.2	Results	42
6.4	Application	44
6.4.1	Exploratory Analysis	45
6.4.2	Linear Model with PAR Errors	46
6.5	Conclusions	51
6.6	Acknowledgements	52
7	CONCLUSÕES E TRABALHOS FUTUROS	53
	REFERÊNCIAS BIBLIOGRÁFICAS	54

1 Introdução

A qualidade do ar de uma região é determinada pela proximidade de fontes poluidoras e pelos níveis de emissão dos poluentes das mesmas, pela capacidade com que a atmosfera da região consegue absorver, dispersar e remover esses contaminantes. Parâmetros meteorológicos como temperatura, pressão atmosférica, umidade relativa do ar, velocidade e direção predominante do vento podem ampliar ou reduzir a capacidade do transporte e de dispersão atmosférica dos poluentes.

As emissões de poluentes atmosféricos podem ser classificadas em antropogênicas e naturais (GODISH, 1997). As emissões antropogênicas são aquelas provocadas pela ação do homem, geralmente nas indústrias, nos transportes e em processos de geração de energia. As emissões naturais são causadas por processos naturais, tais como queimadas naturais, partículas do solo ressuspensas pelo vento, etc. Quanto à origem, os poluentes são classificados em níveis primários ou secundários. Os poluentes primários são aqueles lançados diretamente na atmosfera, como resultado dos processos industriais, dos gases de exaustão dos motores de combustão interna, construção civil e outros. Os poluentes secundários são aqueles formados a partir de reações químicas que ocorrem na atmosfera entre os poluentes primários.

O material particulado (MP) é considerado um poluente atmosférico pelo efeito danoso à saúde das pessoas, dos animais e da vegetação, pela interferência nas mudanças climáticas regionais e globais, e pelo incômodo de sua deposição nas superfícies dos materiais e edificações (WHO, 2005; JACOBSON, 2002). A gravidade dos danos causados pelo MP levou à uma maior abordagem sobre o assunto na literatura e ao crescimento do monitoramento de fontes de emissões e de dados da qualidade do ar em várias regiões do mundo.

O material particulado é composto de partículas capazes de permanecer em suspensão na atmosfera devido às suas pequenas dimensões. Como exemplos podem ser citados a poeira, a fuligem e as partículas de óleo (BRAGA et al., 2005). A forma e a composição química do MP podem ser bastante diversificadas. Normalmente a classificação é feita de acordo com o tamanho da partícula: nos casos de diâmetros aerodinâmicos inferiores a $2,5\mu\text{m}$ e $10\mu\text{m}$ são denominados $\text{MP}_{2,5}$ e MP_{10} , respectivamente. Na literatura, o MP_{10} também é definido como partículas inaláveis. Segundo Baird (2001), as partículas são classificadas como grossas (diâmetro maior que $2,5\mu\text{m}$) e finas (diâmetro menor que $2,5\mu\text{m}$). A importância do tamanho das partículas está relacionada aos danos que elas podem causar à saúde. Holgate et al. (1999) afirmam que as partículas finas são as principais responsáveis por esses danos, uma vez que podem atingir e prejudicar o sistema respiratório inferior.

Tendo em vista os efeitos do material particulado na saúde e no meio ambiente, com o passar do tempo, têm surgido legislações ambientais para regulamentar os níveis de emissões e de qualidade do ar. A legislação brasileira, através da resolução CONAMA nº 3 de 1990, estabeleceu os seguintes padrões primário e secundário de concentração de partículas inaláveis: (1) a concentração média aritmética anual deve ser de, no máximo, $50 \mu\text{g}/\text{m}^3$; e (2) a concentração média de 24 horas deve ser de, no máximo, $150 \mu\text{g}/\text{m}^3$, não podendo ser ultrapassada mais de uma vez por ano. A Organização Mundial de Saúde (OMS) estabeleceu, no ano de 2005, as diretrizes de $50 \mu\text{g}/\text{m}^3$ para a concentração média de 24 horas e de $20 \mu\text{g}/\text{m}^3$ para a média aritmética anual de MP_{10} (WHO, 2005).

Para auxiliar a manutenção dos níveis de concentrações dentro dos padrões da legislação, é primordial a realização do monitoramento da qualidade do ar. Na região metropolitana da Grande Vitória (RMGV), a Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr), é responsável por desempenhar esse papel. A rede foi inaugurada em julho de 2000 pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA) e fornece dados horários de diversos poluentes, sendo que o poluente MP_{10} é monitorado por todas as estações que compõem a RAMQAr.

Esta disponibilidade de dados monitorados sequencialmente no tempo, aliada à necessidade de avaliar a qualidade do ar e de fazer previsões de concentrações desses poluentes, justificam a abordagem de séries temporais para os dados de poluição atmosférica coletados pela RAMQAr. A escolha da RMGV se justifica pelos seguintes fatos sobre a mesma: possui diversas fontes de emissão de poluentes atmosféricos devido à suas indústrias e ao seu crescente desenvolvimento urbano; no que se refere ao efeito dos poluentes na saúde da população, houve um aumento do número de atendimentos hospitalares por doenças respiratórias e cardiovasculares em função de seu crescimento (SOUZA et al., 2014).

Na prática, são utilizados modelos para se entender a dinâmica e prever valores futuros de séries temporais. Usualmente, os modelos mais difundidos na literatura partem do princípio que a mesma é estacionária. Contudo, essa hipótese pode ser muito restritiva, uma vez que existem muitos fenômenos que causam não estacionariedade, tais como: tendências determinísticas, heterocedasticidade, mudanças de nível, raízes unitárias, entre outros. Entre esses fenômenos está a correlação periódica (PC). Processos com correlação periódica, ou periodicamente estacionários (PS) foram introduzidos por Gladyshev (1961). A PC requer atenção especial devido ao fato de não ser revelada por técnicas estacionárias usuais apropriadas a dados estacionários, o que pode levar a especificação incorreta (TIAO; GRUPE, 1980).

Um dos modelos para séries temporais PC é o modelo periódico autorregressivo (PAR), que é uma generalização do modelo autorregressivo proposto por Box & Jenkins (1970), no qual os coeficientes podem variar periodicamente com o tempo. Assim sendo, a estrutura de correlação da série varia periodicamente também e a mesma se adequa melhor às mudanças que ocorrem nas séries diárias de poluentes atmosféricos na RMGV, principal-

mente nos finais de semana.

Outro problema típico das séries temporais de poluentes atmosféricos na RMGV é a presença de picos na série que apesar de não poderem ser considerados observações atípicas por causa da alta frequência com a qual ocorrem na série, podem ser modelados como observações atípicas aditivas já que provocam um efeito na série similar ao causado por tal tipo de observação, isto é, têm efeito de observações que fogem do padrão da série e que não têm nenhum efeito nas observações subsequentes (FOX, 1972). O trabalho de Sarnaglia, Reisen & Lévy-Leduc (2010) mostra como a presença de observações atípicas aditivas degenera a estrutura de autocorrelação de processos periódicos.

Neste estudo, serão utilizadas médias diárias do MP_{10} para garantir similaridade com o disposto na resolução CONAMA n° 3 de 1990. Fisicamente, a ocorrência de PC, ou cicloestacionariedade, em médias diárias de MP_{10} , pode ser justificada por mudanças na intensidade em que fontes poluidoras emitem esse poluente ao longo dos dias da semana. Já a presença de picos de concentração se deve a fugas diárias dos padrões de emissões de fonte e/ou meteorológicos. Os picos causam um problema que é recorrente na área de poluição do ar: deteriorações da qualidade do ar de uma região nos intervalos de tempo em que os mesmos ocorrem. Adicionalmente, esses picos podem prejudicar a eficiência dos estimadores do modelo PAR, como mencionado previamente.

A Figura 1 mostra diversos gráficos da série temporal em estudo depois de uma transformação estabilizadora de variância ($\log(MP_{10})$). São mostrados os picos que ela possui (gráfico de $\log(MP_{10})$), sua mudança de nível nos finais de semana (box-plot periódico de $\log(MP_{10})$), além da mudança da autocorrelação existente entre dias da semana adjacentes (diagramas de dispersão de $\log(MP_{10})$ para terça e quarta além de quarta e quinta). Esses gráficos mostram que tal série é contaminada por outliers aditivos e possui funções média e de autocorrelação periódicas, sendo assim necessário o ajuste de modelos periódicos robustos à mesma.

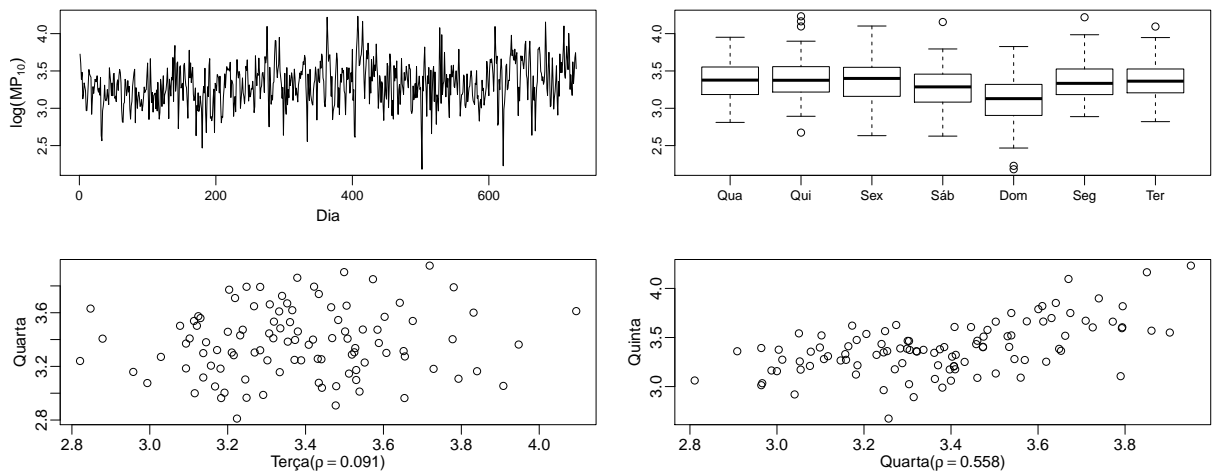


Figura 1 – Série temporal em estudo após transformação logarítmica com box-plot e gráficos de dispersão em defasagem unitária para diferentes dias da semana.

Para isso, será feito inicialmente, o estudo empírico de estimadores robustos do modelo PAR para se saber qual metodologia apresenta propriedades desejadas para a aplicação aos dados de poluição do ar. Neste estudo serão comparados os estimadores propostos por Sarnaglia, Reisen & Lévy-Leduc (2010) (Yule-Walker Robusto) e Shao (2008) (Mínimos Quadrados Robusto) em diversos cenários, tais como: com e sem observações atípicas, além de na presença e ausência de normalidade. Além disso, como forma de comprovar que as metodologias robustas são mais eficientes na presença de observações atípicas, também será estudada metodologia clássica de Yule-Walker (MCLEOD, 1994). Assim, se poderá determinar qual metodologia é mais vantajosa em cada cenário. Posteriormente, todas as três abordagens serão aplicadas em dados de MP_{10} coletados pela estação da RAMQAr da Enseada do Suá.

O uso de modelos periódicos na área de poluição do ar é um dos fatores que motivam o desenvolvimento desta pesquisa. Exemplos de trabalhos que ajustaram modelos periódicos a séries de poluentes atmosféricos são Bloomfield, Hurd & Lund (1994), Sarnaglia, Reisen & Bondon (2015) e Sarnaglia et al. (2016). Os trabalhos das metodologias robustas de estimação de modelos periódicos que serão estudados nesta dissertação não utilizaram dados de poluição do ar na parte de aplicação, essa é mais uma motivação para fazer a aplicação das mesmas aos dados de MP_{10} coletados pela estação da RAMQAr da Enseada do Suá.

A outra motivação é de verificar empiricamente em que cenários a hipótese de simetria dos dados que é requerida na teoria das duas metodologias robustas realmente se faz necessária, uma vez que dados de poluição do ar em geral são assimétricos. Além disso, também é interessante verificar que vantagens a metodologia clássica tem sobre as robustas em um cenário sem contaminação e que vantagens as metodologias robustas têm sobre a clássica em um cenário com contaminação por observações atípicas aditivas.

A justificativa de se utilizar a estação da Enseada do Suá é que esta é uma área que tem como uma das principais fontes do poluente MP_{10} , o intenso tráfego de veículos da região. Tal fonte sofre alteração de acordo com o dia da semana, principalmente, nos finais de semana, nos quais o fluxo de veículos é significativamente menor e o MP_{10} na região tem uma redução do seu valor médio.

O restante desta dissertação está dividido da seguinte forma: no próximo capítulo são apresentados o objetivo geral e os específicos; no capítulo 3 é feita uma revisão da literatura referente à estimação de modelos de séries temporais periódicos autorregressivos robustos; no capítulo 4 são expostos conceitos estatísticos fundamentais; no capítulo 5 são exibidos os materiais e métodos utilizados; no capítulo 6 são mostrados os resultados desta dissertação compilados no formato de um artigo, que incluem a teoria e os resultados; no capítulo 7 são elaboradas as conclusões.

2 Objetivos

2.1 Objetivo Geral

O objetivo geral desta dissertação é estudar a estimação por diversas metodologias de modelos periódicos autorregressivos (PAR) em dados de MP_{10} da Enseada do Suá no intuito de saber qual das metodologias consideradas se adéqua melhor à não gaussianidade, assimetria e presença de observações atípicas dos dados.

2.2 Objetivos Específicos

Os objetivos específicos desta dissertação são os seguintes:

- Estudar a estimação do modelo PAR na presença de distribuição não gaussiana assimétrica e de observações atípicas;
- Investigar empiricamente modelos clássicos e robustos de estimação de processos periodicamente estacionários;
- Utilizar os métodos considerados na estimação de modelos PAR para as séries de MP_{10} da Enseada do Suá.

3 Revisão Bibliográfica

O estudo de séries temporais geralmente se baseia na hipótese de estacionariedade. Existem várias referências clássicas na análise de séries temporais estacionárias, tais como Wei et al. (2006), Brockwell & Davis (2006) e Priestley (1981). No entanto, fenômenos não estacionários são muito comuns e como exemplos de tais processos pode-se citar: raízes unitárias, tendências determinísticas, heterocedasticidade, entre outros.

Outro exemplo de processos não estacionários são aqueles que possuem covariâncias variando periodicamente, tais processos têm sido denominados periodicamente correlacionados (PC) (também conhecidos como periodicamente estacionários ou cicloestacionários) e foram introduzidos por Gladyshev (1961).

A ocorrência de processos periodicamente correlacionados é corroborada por aplicações reais em muitas áreas, por exemplo, (GARDNER; FRANKS, 1975) investigaram a cicloestacionariedade em engenharia elétrica e Bloomfield, Hurd & Lund (1994) estudaram dados de ozônio estratosférico. Uma revisão mais recente em processos PC pode ser vista em Gardner, Napolitano & Paura (2006). Além disso, vale destacar aplicações em outras disciplinas como economia (FRANSES; PAAP, 2004) e hidrologia (MCLEOD, 1994).

No caso de dados com características sazonais, uma prática comum na literatura consiste em ajustar aos dados um modelo sazonal depois de remover qualquer tendência. Como é mostrado em Vecchia & Ballerini (1991), se a estrutura de covariância tem um comportamento não estacionário periódico (ou cíclico), esse tipo de estratégia é inapropriada e deteriora a performance de previsão. Essa especificação incorreta não é revelada pela usual checagem de diagnóstico residual (GARDNER; FRANKS, 1975).

Existem várias maneiras de modelar e prever séries temporais sazonais que não foram sazonalmente ajustadas, como pode ser visto em Franses & Paap (2002). São elas:

- Baseando-se na metodologia de Box & Jenkins (1970); onde são considerados os modelos Sazonais Autorregressivos Integrados de Médias Móveis (SARIMA);
- Assumindo que a série temporal pode ser decomposta em tendência, ciclo, sazonalidade e erros aleatórios que podem seguir os modelos SARIMA, por exemplo;
- Assumindo que a variedade sazonal é melhor descrita permitindo que os parâmetros da autorregressão variem com as estações periódicas, formando assim, o modelo PAR;
- Levando-se em conta um modelo autorregressivo de médias móveis no qual os parâmetros podem variar com as estações periódicas (modelo PARMA).

Inicialmente, os processos autorregressivos periódicos foram tratados como se pode ver em Jones & Brelsford (1967), que foca no problema de prever séries temporais com estrutura periódica, incluindo a estimação dos parâmetros necessários e das variâncias; Troutman (1979) que mostra, entre outros assuntos, a representação do processo em filtro linear e as restrições aos parâmetros dos modelos que devem ser satisfeitas para permitir tal representação e McLeod (1994) que trata da checagem de diagnóstico de modelos periódicos autorregressivos. Para uma revisão mais completa sobre séries temporais periodicamente correlacionadas recomenda-se a leitura de Hurd & Miamer (2007).

O jeito mais fácil de construir modelos para processos periodicamente correlacionados é permitir que parâmetros de modelos estacionários variem com o tempo. Nesse contexto, o modelo PAR surge como uma extensão do modelo autorregressivo amplamente divulgado na literatura.

Surpreendentemente, apesar do fato do modelo Periódico Autorregressivo Integrado de Médias Móveis (PARMA) ser mais parcimonioso, ele ainda não foi muito utilizado em aplicações reais. As maiores razões para isso podem ser por dificuldade de implementação de métodos de estimação e seus esforços computacionais. Dessa forma, em muitos casos continua sendo mais vantajoso o uso de modelos PAR.

A previsão de modelos PAR é feita basicamente da mesma maneira que a previsão de modelos AR. Para modelos com ordens maiores que um, pode ser conveniente utilizar a representação vetorial de processos PAR para computar previsões e variâncias de erros de previsões, veja Franses & Paap (2002) para maiores detalhes. Nesse caso, revisões podem ser geradas de uma maneira semelhante à de modelos Autorregressivos Vetoriais (VAR), como em Lütkepohl & Poskitt (1991).

Para descobrir como se identifica a presença de periodicidade sugere-se inicialmente analisar os diagramas de dispersão apropriados conforme mostra McLeod (1994). Em casos que não seja possível determinar a presença de periodicidade através de tais gráficos, as metodologias sugeridas por Hurd & Gerr (1991), Vecchia & Ballerini (1991), Bloomfield, Hurd & Lund (1994) e Broszkiewicz-Suwaj et al. (2004) podem ser utilizadas.

Além disso existem metodologias que são utilizadas exclusivamente para se descobrir o período sazonal de uma série, após a periodicidade das mesmas já ter sido detectada. A principal motivação para isso é que quando os períodos das funções média e de autocovariância são diferentes, alguns métodos podem falhar em estimar corretamente o período da série. Assim sendo, é mais adequada a utilização de metodologias como as de Martin & Kedem (1993) e Wang, Chen & Huang (2006).

Ao considerar um modelo $PAR(p)$, a ordem p pode ser determinada utilizando o gráfico da função de autocorrelação parcial (FACP) amostral periódica ou através de critérios de informação como o Critério de Informação de Akaike (AIC) periódico ou o Critério de Informação de Schwartz (BIC) periódico (MCLEOD, 1994). Ambos os critérios supracitados podem ser implementados usando os métodos desenvolvidos por

Sakai (1982).

A estimação de parâmetros do modelo PAR já é bem documentada na literatura. Exemplos de estimadores de modelos PAR são o de Yule-Walker (MCLEOD, 1994), o de Máxima Verossimilhança (LUND; BASAWA, 2000) e o de Mínimos Quadrados (BASAWA; LUND, 2001). No entanto, na maioria das aplicações há a presença de picos que, apesar de não o serem, se comportam como observações que fogem do padrão da amostra (observações atípicas) e a adoção de metodologias clássicas pode subestimar os valores dos parâmetros do modelo.

Dessa forma, se faz necessária a adoção de metodologias robustas que diminuam o impacto adverso causado por esses picos na série, sem a necessidade de retirá-los da amostra. Como pode ser visto em Fox (1972), existem diversos tipos de observações atípicas, mas nesta dissertação serão abordados apenas os do tipo aditivo uma vez que elas são as que ocorrem com maior frequência nos dados de MP_{10} da estação da RAMQAr da Enseada do Suá. Para uma revisão mais completa sobre estatística robusta, recomenda-se a leitura de Maronna, Martin & Yohai (2006).

Várias metodologias podem ser adotadas para se estimar um modelo PAR robusto. Shao (2008) propôs a utilização de uma generalização do M-estimador com a função ψ de Huber (HUBER et al., 1964) conjuntamente com o estimador de dispersão MAD para a estimação dos erros padronizados de um modelo e a utilização da M-regressão que pode ser vista em Maronna, Martin & Yohai (2006), por exemplo. Já Sarnaglia, Reisen & Lévy-Leduc (2010) propuseram generalizações para o estimador de escala robusto Q_n proposto por Rousseeuw & Croux (1993) e para o estimador da função de autocovariância robusta proposto por Ma & Genton (2000). Essas funções periódicas robustas foram usadas nas equações de Yule-Walker para obter as estimativas robustas de parâmetros.

Nesta dissertação, serão comparadas as metodologias de estimação periódica clássicas de Yule-Walker (MCLEOD, 1994) com as metodologias robustas de Yule-Walker (SARNAGLIA; REISEN; LÉVY-LEDUC, 2010) e de Mínimos Quadrados (SHAO, 2008) em simulação e aplicação. O objetivo é avaliar empiricamente qual efeito a assimetria e a presença de picos nos dados do poluente MP_{10} da estação da RAMQAr da Enseada do Suá pode ter nos estimadores estudados nesta dissertação, de forma a saber qual metodologia é a mais adequada para a aplicação à esses dados.

4 Conceitos Fundamentais

Nesta seção serão apresentados alguns conceitos estatísticos fundamentais para a discussão apresentada nesta dissertação. Para o desenvolvimento mais aprofundado dos assuntos abordados neste capítulo, recomenda-se a leitura de Box & Jenkins (1970), Brockwell & Davis (2006), Priestley (1981) e Wei et al. (2006).

4.1 Processo Estocástico

Definição 1. *Um processo estocástico é uma família de variáveis aleatórias indexadas no tempo $\{Y_t(\omega), \omega \in \Omega, t \in \tau\}$, onde Ω denota o espaço amostral e τ é um conjunto de índices arbitrário.*

Nesta dissertação, $\tau = \mathbb{Z}$ o conjunto dos inteiros e o índice t representa o tempo. Fixado t , $Y_t(\omega)$ é uma variável aleatória e para um dado ω , a sequência numérica $\{Y_t(\omega), t \in \mathbb{Z}\}$ é denominada trajetória do processo estocástico. Uma trajetória finita $\{Y_t(\omega), t = 1, \dots, n\}$ é denominada série temporal. De agora em diante, o argumento ω será omitido e o processo estocástico será representado simplesmente por $\{Y_t\}$.

A função média de um processo estocástico $\{Y_t\}$ é definida por

$$\mu_t = E(Y_t).$$

Já a função autocovariância (ACV) do processo $\{Y_t\}$ é definida por

$$\gamma(t_1, t_2) = \text{Cov}(Y_{t_1}, Y_{t_2}) = E[(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2})].$$

Em particular, a função variância é $\gamma_0(t) = \gamma(t, t)$. A função autocorrelação (ACF) é

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sqrt{\sigma_{t_1}^2} \sqrt{\sigma_{t_2}^2}} = \frac{E[(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2})]}{\sqrt{E[(Y_{t_1} - \mu_{t_1})^2]} \sqrt{E[(Y_{t_2} - \mu_{t_2})^2]}}.$$

4.2 Processos Estocásticos Estacionários

Em geral, para realizar inferência com respeito a um processo estocástico, é necessário que se exija estabilidade do mesmo. Nesse sentido, definem-se diversos tipos de estacionariedade. Especificamente, considere uma subsequência finita $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$ de um processo $\{Y_t\}$. A função de distribuição de probabilidade acumulada conjunta da subsequência no ponto $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ é definida por

$$F_{Y_{t_1}, \dots, Y_{t_n}}(x_1, \dots, x_n) = P\{Y_{t_1} \leq x_1, \dots, Y_{t_n} \leq x_n\}.$$

Um processo é dito ser estritamente estacionário de primeira ordem em distribuição se sua função de distribuição de probabilidade unidimensional não varia sob translações do tempo, isto é, se $F_{Y_{t_1}}(x_1) = F_{Y_{t_1+h}}(x_1)$ para quaisquer inteiros t_1 e h ; estritamente estacionário de segunda ordem se $F_{Y_{t_1}, Y_{t_2}}(x_1, x_2) = F_{Y_{t_1+h}, Y_{t_2+h}}(x_1, x_2)$ para quaisquer inteiros t_1, t_2 e h . De forma mais geral tem-se a seguinte definição de processos estacionários.

Definição 2. *Um processo é estritamente estacionário de ordem n se*

$$F_{Y_{t_1}, \dots, Y_{t_n}}(x_1, \dots, x_n) = F_{Y_{t_1+h}, \dots, Y_{t_n+h}}(x_1, \dots, x_n)$$

para quaisquer $t_1, \dots, t_n, h \in \mathbb{Z}$.

Um processo é dito ser estritamente (ou fortemente) estacionário se a Definição 2 é verdadeira qualquer $n \in \mathbb{Z}$. Um processo estritamente estacionário de ordem n também é estritamente estacionário de ordem m , desde que $m \leq n$, pois a função de distribuição de n -ésima ordem determina todas as funções de distribuição de ordem inferior.

Usualmente, a estacionariedade estrita é relaxada, e são considerados processos fracamente estacionários. Em particular, os processos fracamente estacionários de 2ª ordem desempenham papel fundamental na literatura de análise de séries temporais.

Definição 3. *Um processo $\{Y_t\}$ é dito ser fracamente estacionário de 2ª ordem (ou simplesmente estacionário) se $E(Y_t^2) < \infty$, $\mu_t = \mu$ e $\gamma(t_1, t_2) = \gamma(t_1 + h, t_2 + h)$, para todo $t, t_1, t_2, h \in \mathbb{Z}$.*

Pela Definição 3, tomando $h = -t_2$, nota-se que

$$\gamma(t_1, t_2) = \gamma(t_1 + h, t_2 + h) = \gamma(t_1 - t_2, 0) = \gamma_{t_1 - t_2}.$$

Assim, para processos fracamente estacionários, a ACV do processo $\{Y_t\}$ depende apenas da diferença de tempo entre as observações.

4.3 Funções de Autocovariância e de Autocorrelação

Para um processo estacionário $\{Y_t\}$, a média $E(Y_t) = \mu$ e a variância $\text{Var}(Y_t) = \gamma(t, t) = \gamma_0$ são constantes e a ACV $\text{Cov}(Y_t, Y_s)$ é função apenas da diferença de tempo $t - s$. Assim, nesse caso, pode-se escrever a ACV entre Y_t e Y_{t+h} como

$$\gamma_h = \text{Cov}(Y_t, Y_{t+h}) = E[(Y_t - \mu)(Y_{t+h} - \mu)],$$

e a ACF entre Y_t e Y_{t+h} como

$$\rho_h = \frac{\text{Cov}(Y_t, Y_{t+h})}{\sqrt{\text{Var}(Y_t)}\sqrt{\text{Var}(Y_{t+h})}} = \frac{\gamma_h}{\gamma_0}.$$

O termo h é denominado defasagem ou *lag*.

É fácil ver que, para um processo estacionário, a função de autocovariância γ_h e a função de autocorrelação ρ_h têm as seguintes propriedades:

1. $\gamma_0 = \text{Var}(Y_t)$; $\rho_0 = 1$.
2. $|\gamma_h| \leq \gamma_0$; $|\rho_h| \leq 1$.
3. $\gamma_h = \gamma_{-h}$ e $\rho_h = \rho_{-h}$ para todo $h \in \mathbb{Z}$, isto é, γ_h e ρ_h são funções pares e portanto simétricas em torno de $h = 0$. Assim, a função de autocorrelação é frequentemente plotada apenas para defasagens não negativas.
4. Outra propriedade importante da função de autocovariância γ_h e de autocorrelação ρ_h é que elas são funções positivas semidefinidas, isto é,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \gamma_{|t_i - t_j|} \geq 0$$

e

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho_{|t_i - t_j|} \geq 0$$

para quaisquer $t_1, t_2, \dots, t_n \in \mathbb{Z}$ e quaisquer $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$.

4.4 Função de Autocorrelação Parcial

Além da autocorreção entre Y_t e Y_{t+h} , pode-se querer investigar a correlação entre Y_t e Y_{t+h} depois da dependência linear mútua das variáveis intermediárias $Y_{t+1}, Y_{t+2}, \dots, Y_{t+h-1}$ ser removida. Assim, define-se a função de autocorrelação parcial (FACP) de $\{Y_t\}$ para a defasagem h como

$$\phi_{hh} = \text{Corr}(Y_t, Y_{t+h} | Y_{t+1}, \dots, Y_{t+h-1}).$$

4.5 Processos Ruído Branco

Um processo $\{\epsilon_t\}$ pode ser considerado ruído branco se for uma sequência de variáveis aleatórias não correlacionadas com média constante $E(\epsilon_t) = \mu_\epsilon$ (que geralmente é assumida como 0) e variância também constante $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$. Segue-se por definição que um processo ruído branco é estacionário com função de autocovariância

$$\gamma_h = \begin{cases} \sigma_\epsilon^2, & h = 0, \\ 0, & h \neq 0, \end{cases}$$

função de autocorrelação

$$\rho_h = \begin{cases} 1, & h = 0, \\ 0, & h \neq 0, \end{cases}$$

e função de autocorrelação parcial

$$\phi_{hh} = \begin{cases} 1, & h = 0, \\ 0, & h \neq 0. \end{cases}$$

4.6 Modelos Estacionários de Séries Temporais

4.6.1 Representações de Médias Móveis e Autorregressiva de Modelos de Séries Temporais

Em análise de séries temporais, existem duas representações úteis para modelagem. Uma é escrever o desvio de um processo com relação à sua média, $\{\dot{Y}_t\} = \{Y_t - \mu\}$, como uma combinação linear de uma sequência de variáveis aleatórias não correlacionadas, isto é,

$$\dot{Y}_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \cdots = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \quad (4.1)$$

onde $\psi_0 = 1$, $\{\epsilon_t\}$ é um processo ruído branco de média zero e $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. Introduzindo o operador retardo $B^m \dot{Y}_t = \dot{Y}_{t-m}$, pode-se escrever (4.1) na forma compacta

$$\dot{Y}_t = \psi(B) \epsilon_t,$$

onde $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

A forma de (4.1) é chamada de representação de médias móveis (MA) de ordem infinita de um processo. Wold (1938) provou que um processo estacionário puramente não determinístico sempre pode ser expresso na forma (4.1).

Outra forma útil de se escrever $\{\dot{Y}_t\}$, é utilizando a representação autorregressiva (AR) de ordem infinita, na qual o valor de \dot{Y}_t é representado através de uma equação de regressão em termos dos seus próprios valores passados mais um choque aleatório, isto é,

$$\dot{Y}_t = \pi_1 \dot{Y}_{t-1} + \pi_2 \dot{Y}_{t-2} + \cdots + \epsilon_t. \quad (4.2)$$

ou, equivalentemente,

$$\pi(B) \dot{Y}_t = \epsilon_t,$$

onde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ e $1 + \sum_{j=1}^{\infty} |\pi_j| < \infty$. Box & Jenkins (1976) chamaram os processos que podem ser escritos na forma de (4.2) de invertíveis.

4.6.2 Modelo Autorregressivo

Na representação AR de um processo, se apenas um número finito de pesos π são diferentes de zero, isto é, $\pi_1 = \phi_1, \pi_2 = \phi_2, \dots, \pi_p = \phi_p$ e $\pi_k = 0$, para $k > p$, dessa forma, o processo resultante é chamado de processo ou modelo autorregressivo de ordem p e é denotado como $AR(p)$, tal processo pode ser expresso da seguinte forma

$$\dot{Y}_t = \phi_1 \dot{Y}_{t-1} + \cdots + \phi_p \dot{Y}_{t-p} + \epsilon_t.$$

Pode-se escrever o processo $AR(p)$ como

$$\phi_p(B) \dot{Y}_t = \epsilon_t,$$

onde $\phi_p(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$. O processo AR é sempre invertível. Sua condição de estacionariedade é que as raízes G_1, G_2, \dots, G_p , da equação característica $\phi_p(z) = (1 - G_1^{-1}z)(1 - G_2^{-1}z) \dots (1 - G_p^{-1}z) = 0, z \in \mathbb{C}$, estejam estritamente fora do círculo unitário, isto é, $|G_i| > 1$. Por exemplo, para o processo AR(1) isso equivale a $|\phi_1| < 1$.

4.6.3 Modelo de Médias Móveis

Na representação de médias móveis de um processo, se apenas um número finito de pesos ψ são diferentes de zero, isto é, $\psi_1 = -\theta_1, \psi_2 = -\theta_2, \dots, \psi_q = -\theta_q$ e $\psi_k = 0$ para $k > q$, diz-se que $\{\dot{Y}_t\}$ é um processo MA de ordem q , denotado por MA(q), tal processo pode ser expresso da seguinte forma

$$\dot{Y}_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q}$$

ou

$$\dot{Y}_t = \theta_q(B) \epsilon_t,$$

onde $\theta_q(B) = (1 - \theta_1 B - \dots - \theta_q B^q)$.

O processo MA é sempre estacionário. Sua condição de invertibilidade é que as raízes H_1, H_2, \dots, H_q , da equação característica $\theta_q(z) = (1 - H_1^{-1}z)(1 - H_2^{-1}z) \dots (1 - H_q^{-1}z) = 0, z \in \mathbb{C}$, estejam estritamente fora do círculo unitário, isto é, $|H_i| > 1$. Por exemplo, para o processo MA(1) isso equivale a $|\theta_1| < 1$.

4.6.4 Modelo Autorregressivo de Médias Móveis

Uma extensão natural das representações AR pura e MA pura é a representação mista autorregressiva de médias móveis (ARMA), que inclui as representações AR e MA como casos especiais. Essa representação se faz necessária para se obter um modelo mais parcimonioso (com menos parâmetros) em casos que as representações AR ou de MA necessitam de ordens muito grandes para boa aproximação. Para um modelo ARMA(p, q) de ordem autorregressiva p e ordem de médias móveis q temos

$$(1 - \phi_1 B - \dots - \phi_p B^p) \dot{Y}_t = (1 - \theta_1 B - \dots - \theta_q B^q) \epsilon_t$$

ou

$$\phi_p(B) \dot{Y}_t = \theta_q(B) \epsilon_t.$$

A condição de estacionariedade é que as raízes de $\phi_p(z) = 0, z \in \mathbb{C}$ estejam fora do círculo unitário. Já a condição de invertibilidade é que as raízes de $\theta_q(z) = 0, z \in \mathbb{C}$ estejam fora do círculo unitário. Além disso, os polinômios $\phi_p(z)$ e $\theta_q(z)$ não têm fatores comuns.

4.7 Sazonalidade

O fenômeno sazonalidade é caracterizado por picos de autocorrelação nas defasagens sazonais, o que indica alta correlação entre observações no mesmo período (*season*). Formalmente, representando o período sazonal do processo $\{\dot{Y}_t\}$ por s , a sazonalidade induz alta correlação entre as observações \dot{Y}_t e \dot{Y}_{t+ks} , principalmente se k é pequeno.

No intuito de explicar os princípios de como um processo sazonal funciona, suponha que o processo é observado em frequência mensal e considere a subsequência constituída apenas das observações de janeiro. Por simplicidade, assuma que o seguinte modelo ARIMA descreva as observações de janeiro

$$\Phi_P(L)(1-L)^D \dot{Y}_t = \Theta_Q(L)a_t, \quad (4.3)$$

onde $\Phi_P(z) = (1 - \Phi_1 z - \dots - \Phi_P z^P)$ denota o polinômio AR e $\Theta_Q(z) = (1 - \Theta_1 z - \dots - \Theta_Q z^Q)$ representa o polinômio MA. O operador retardo sazonal $L = B^{12}$, serve para o mesmo propósito que o operador retardo B , mas retarda o tempo em 12 unidades, para uma série com sazonalidade s , temos $L = B^s$. Os erros $a_t, a_{t-12}, a_{t-24}, \dots$ são não correlacionados.

Suponha ainda que o mesmo modelo vale para os outros meses: fevereiro, março, ..., dezembro. Isso implica que todos os erros correspondentes a um mês fixo em anos diferentes são não correlacionados. Porém, em situações práticas, não é intuitivo assumir que erros correspondentes a meses adjacentes sejam não correlacionados. Em outras palavras a série “residual” $a_t, a_{t-1}, a_{t-2}, \dots$ pode ser correlacionada. É comum modelar essa estrutura de correlação remanescente através de um modelo ARIMA, isto é, assume-se que $\{a_t\}$ satisfaz

$$\phi_p(B)(1-B)^d a_t = \theta_q(B)\epsilon_t, \quad (4.4)$$

onde $\phi_p(z) = (1 - \phi_1 z - \dots - \phi_p z^p)$ denota o polinômio AR e $\theta_q(z) = (1 - \theta_1 z - \dots - \theta_q z^q)$ representa o polinômio MA, além disso, $\{\epsilon_t\}$ é uma sequência ruído branco. Assim, combinando os modelos (4.3) e (4.4), e tomando $L = B^s$, obtém-se o modelo sazonal autorregressivo integrado de médias móveis de ordens não sazonais p, d, q e sazonais P, D, Q e período sazonal s (SARIMA(p, d, q) \times (P, D, Q) $_s$)

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D \dot{Y}_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t.$$

4.8 Modelos Não Estacionários de Séries Temporais

Na prática, a hipótese de estacionariedade fraca pode ser demasiadamente restritiva, uma vez que não são raros os casos em que o conjunto de dados apresenta média e/ou variância não constantes. No primeiro caso, modelos de tendência determinística ou diferenciação podem ser capazes de corrigir esse problema. No segundo caso, comumente, utilizam-se transformações estabilizadoras de variância.

4.8.1 Modelos de Tendência Determinísticos

Em geral, existem casos em que a função média de um processo não estacionário pode ser representada por uma função determinística do tempo. Nesse cenário, modelos de regressão podem ser utilizados para descrever o fenômeno. Isto é, o processo pode ser escrito como $Y_t = \mu_t + \epsilon_t$, onde $\{\epsilon_t\}$ é um processo estacionário de média zero e μ_t descreve o comportamento funcional da função média. Por exemplo, se é razoável assumir que a função média μ_t se comporta no tempo como um polinômio de ordem k , isto é, $\mu_t = \alpha_0 + \alpha_1 t + \dots + \alpha_k t^k$, o processo pode ser modelado por

$$Y_t = \alpha_0 + \alpha_1 t + \dots + \alpha_k t^k + \epsilon_t.$$

Se a tendência determinística puder ser representada por uma curva na forma de senoide, pode-se utilizar

$$Y_t = \nu_0 + \nu \cos(\omega t + \theta) + \epsilon_t. \quad (4.5)$$

Equivalentemente, a Equação (4.5) pode ser reescrita como

$$Y_t = \alpha \cos(\omega t) + \beta \sin(\omega t) + \epsilon_t,$$

onde

$$\alpha = \nu \cos \theta, \quad \beta = -\nu \sin \theta, \\ \nu = \sqrt{\alpha^2 + \beta^2}.$$

Nesse caso, os parâmetros ν , ω e θ são denominados amplitude, frequência e fase da curva, respectivamente. Mais genericamente, pode-se ter

$$Y_t = \nu_0 + \sum_{j=1}^m (\alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t)) + \epsilon_t,$$

que é frequentemente chamado de modelo de periodicidades ocultas.

Nas situações em que o fenômeno apresenta características sazonais, é comum utilizar variáveis *dummies* sazonais para descrever a média. Formalmente, o processo $\{Y_t\}$ é escrito como

$$Y_t = \sum_{i=1}^s \alpha_i \delta_{ti} + \epsilon_t,$$

onde $\delta_{ti} = 1$, se t corresponde ao período sazonal i , e $\delta_{ti} = 0$, caso contrário; e α_i corresponde ao nível da série no período sazonal i .

4.8.2 Modelos Autoregressivos Integrados de Médias Móveis

Em alguns casos, a diferenciação pode induzir a estacionariedade, isto é, $\{\dot{Y}_t\}$ é não estacionário, mas $\{(1-B)^d \dot{Y}_t\}_{d \in \mathbb{N}^*}$ o é. Comumente, assume-se que o processo diferenciado

$\{(1 - B)^d \dot{Y}_t\}$ segue um modelo ARMA(p, q), o que, conseqüentemente, produz o seguinte modelo para o processo $\{\dot{Y}_t\}$

$$\phi_p(B)(1 - B)^d \dot{Y}_t = \theta_q(B)\epsilon_t,$$

onde $\phi_p(z) = (1 - \phi_1 z - \dots - \phi_p z^p)$, $z \in \mathbb{C}$, é o polinômio AR, $\theta_q(z) = (1 - \theta_1 z - \dots - \theta_q z^q)$, $z \in \mathbb{C}$, é o polinômio MA e $d \in \mathbb{N}^*$ denota a ordem de diferenciação. Nesses casos, dizemos que $\{\dot{Y}_t\}$ segue um modelo Autoregressivo Integrado de Médias Móveis (ARIMA) de ordens p , d e q (ARIMA(p, d, q)).

4.9 Transformações Estabilizadoras da Variância

Nem sempre a não estacionariedade pode ser corrigida através de modelos de tendência determinística ou de diferenciação. Muitas séries temporais podem apresentar não estacionariedade na variância. Para tratar esse problema, pode-se realizar transformações nos dados no intuito de estabilizar a variância.

Nesses casos, é muito comum que a variância do processo estocástico mude quando seu nível muda. Assim, pode-se escrever

$$\text{Var}(Y_t) = cf(\mu_t),$$

para alguma constante $c > 0$ e alguma função f . Nesse contexto, Box & Cox (1964) propuseram a seguinte transformação de potência para corrigir esse problema

$$T(y) = \frac{y^\lambda - 1}{\lambda}. \quad (4.6)$$

Isto é, para algum valor apropriado de λ , $T(Y_t)$ deveria apresentar variância aproximadamente constante, onde $T(\cdot)$ é definida em (4.6). A Tabela 1, mostra alguns dos valores mais comuns usados para λ e suas transformações associadas.

Tabela 1 – Valores comuns de λ e suas transformações associadas.

Valores de λ	Transformação
-1.0	$1/Y_t$
-0.5	$1/\sqrt{Y_t}$
0.0	$\ln Y_t$
0.5	$\sqrt{Y_t}$
1.0	Y_t

Alternativamente, o valor de λ pode ser estimado dos dados utilizando máxima verossimilhança, por exemplo.

4.10 Correlação Periódica

Em alguns casos, a não estacionariedade pode atingir não somente as funções média e variância, mas também as autocorrelações do processo. Nesse contexto, um caso especial ocorre quando a estrutura de autocorrelação é função periódica do tempo.

Vale ressaltar que ferramentas usuais como a ACF e a PACF, em geral, não acusam a presença de correlação periódica, o que pode levar a especificações errôneas (TIAO; GRUPE, 1980). Essa problemática motiva a exploração de técnicas apropriadas a esse tipo de dados. Por esse motivo, processos periódicos são o objeto de estudo desta dissertação e serão discutidos mais detalhadamente no capítulo 6.

É importante ressaltar que a correlação periódica difere da sazonalidade em diversos aspectos. O principal deles é que uma série que apresenta correlação periódica é não estacionária enquanto uma série com sazonalidade o é. Isso se deve ao fato da estrutura de autocorrelação da série mudar para cada período da mesma quando há correlação periódica, já quando a série possui sazonalidade essa estrutura é sempre a mesma.

Além disso, a correlação periódica não necessariamente apresenta altos picos de correlação em defasagens sazonais. Por exemplo, para uma série periodicamente correlacionada observada mensalmente, não necessariamente haverá correlação significativa entre janeiro de um ano e janeiro do ano anterior, porém a intensidade da correlação entre fevereiro e janeiro pode diferir daquela correspondente a março e fevereiro, o que mostra o notável comportamento não estacionário de processos acometidos por este fenômeno. A sazonalidade, por outro lado, necessariamente apresenta correlação significativa em defasagens sazonais.

5 Materiais e Métodos

5.1 Região de Estudo

A Região Metropolitana da Grande Vitória (RMGV) é constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra, Fundão, Guarapari e Viana. A RMGV é localizada na região sudeste do estado do Espírito Santo. Sua área é de 2331,03 km² e possui uma população de aproximadamente 1,884 milhão de habitantes, o que representa cerca de 48% da população total do estado (IBGE, 2011).

A RMGV é o principal polo industrial e econômico do estado, com aproximadamente 63,13% do Produto Interno Bruto (PIB) do Espírito Santo. Nessa região encontram-se atividades de siderurgia, pelletização, mineração (pedreiras), cimenteira, indústria alimentícia, usina de asfalto, etc. No ano de 2007 a RMGV possuía 49,18% da frota veicular total do estado (IJSN, 2008).

No que se refere ao relevo, a RMGV é caracterizada por: cadeias montanhosas nas porções Noroeste (Mestre Álvaro) e Oeste (Região Serrana); planícies (Aeroporto e manguezais) e planaltos (Planalto Serrano) na porção Norte; e planícies (Barra do Jucu) na porção Sul. Todas porções são intercaladas por maciços rochosos de pequeno e médio porte. As condições de relevo no geral são favoráveis, em grande parte da região, à circulação de ventos para dispersão de poluentes (IEMA, 2013).

5.2 Rede Automática de Monitoramento da Qualidade do Ar

O início do funcionamento da Rede Automática de Monitoramento da Qualidade do Ar (RAMQAr) foi no ano de 2000. A rede é de propriedade e responsabilidade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). No período de coleta de dados considerado neste estudo, a rede era composta de 8 estações de monitoramento distribuídas nos municípios da RMGV da seguinte forma: 2 estações no município de Serra (Laranjeiras e Carapina); 3 estações no município de Vitória (Jardim Camburi, Enseada do Suá e Centro); 2 estações no município de Vila Velha (Ibes e Centro) e 1 estação no município de Cariacica (Vila Capixaba). A localização espacial das estações de monitoramento da RAMQAr encontra-se na Figura 2.

Os poluentes monitorados nas 8 estações utilizadas da RAMQAR são: dióxido de enxofre (SO₂), partículas totais em suspensão (PTS), partículas inaláveis (MP₁₀), ozônio (O₃), óxidos de nitrogênio (NO_x), monóxido de carbono (CO) e hidrocarbonetos (HC). Além desses poluentes, alguns parâmetros meteorológicos são monitorados: direção dos ventos (DV), velocidade dos ventos (VV), precipitação pluviométrica (PP), umidade re-

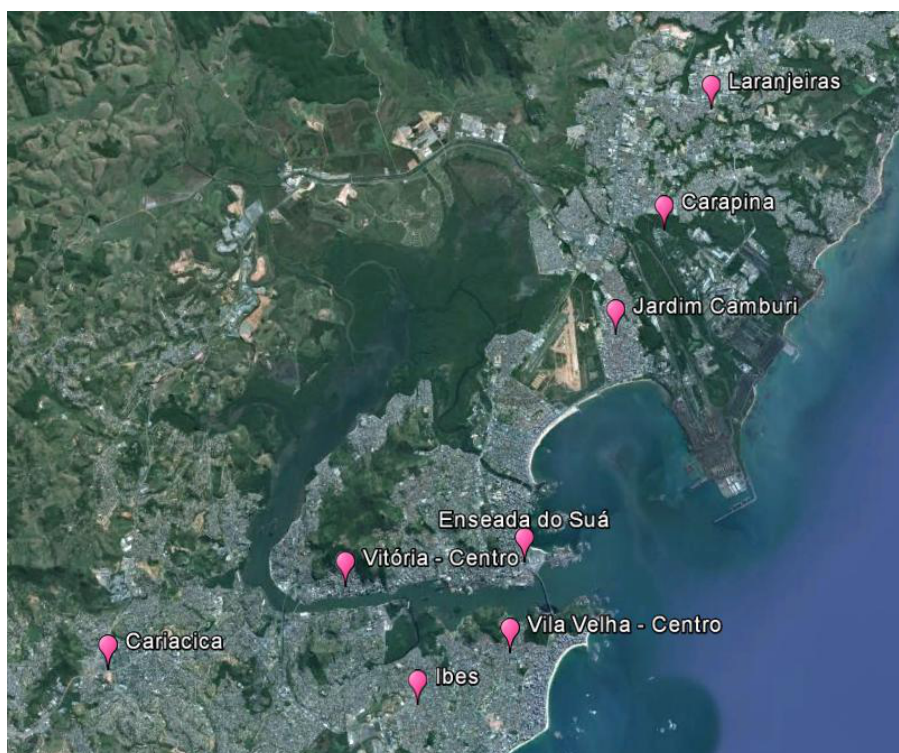


Figura 2 – Localização espacial das estações da RAMQAr.

lativa do ar (UR), temperatura (T), pressão atmosférica (P) e radiação solar (I). Nem todos os poluentes e parâmetros meteorológicos são monitorados por todas as estações. Os poluentes e parâmetros monitorados por cada estação estão mostrados na Tabela 2.

Tabela 2 – Poluentes e parâmetros meteorológicos em cada estação da RAMQAr.

Estação	PTS	MP ₁₀	SO ₂	CO	NO _x	HC	O ₃	Meteorologia
Laranjeiras	X	X	X	X	X		X	
Carapina	X	X						DV,VV,UR,PP,P,T,I
Jardim Camburi	X	X	X		X			
Enseada do Suá	X	X	X	X	X	X	X	DV,VV
Vitória Centro	X	X	X	X	X	X		
Ibes	X	X	X	X	X	X	X	DV,VV
Vila Velha Centro		X	X					
Vila Capixaba	X	X	X	X	X		X	DV,VV,T

A estação Enseada do Suá e a estação Ibes são as únicas que registram as concentrações de todos os poluentes. A estação Carapina monitora todos os parâmetros meteorológicos.

5.3 Dados

Esta dissertação foi realizada utilizando os dados de MP_{10} medidos na estação da RAMQAr da Enseada do Suá. O período de medição foi de janeiro de 2014 a dezembro de 2015. Os dados são fornecidos em médias horárias de concentrações medidas em $\mu\text{g}/\text{m}^3$.

5.4 Software Estatístico

As metodologias consideradas e toda análise efetuada foi realizada no *Software R*. Além de ser gratuito, o R possui um grande número de procedimentos estatísticos convencionais, entre eles estão os modelos lineares, modelos de regressão não linear, análise de séries temporais, testes estatísticos paramétricos e não paramétricos, análise multivariada, etc. Este *software* dispõe ainda de uma grande quantidade de funções para o desenvolvimento de ambiente gráfico e criação de diversos tipos de apresentação de dados (REISEN; SILVA, 2011).

No próximo capítulo são apresentados todos os resultados obtidos nesta pesquisa compilados no formato de um artigo científico que foi submetido ao periódico *Communications in Statistics - Theory and Methods* (B1 no QUALIS na área de Engenharias I).

6 Empirical Study of Robust Estimation Methods for PAR Models with Application to PM_{10} Data

Abstract

This paper presents the study of robust estimation methods in Periodic Autoregressive (PAR) processes, that is, the methods investigated in Sarnaglia, Reisen & Lévy-Leduc (2010) and Shao (2008). The methodology proposed by Shao (2008) (RLSE) seeks robustness through iterative use of ordinary least squares on robust versions of the original time series, while the approach proposed by Sarnaglia, Reisen & Lévy-Leduc (2010) (RYWE) uses a generalization of the robust autocovariance function (MA; GENTON, 2000) in the periodic Yule-Walker equations. For comparison purpose, the classical periodic Yule-Walker estimation (YWE) procedure (MCLEOD, 1994) is considered. The daily mean Particulate Matter concentration (PM_{10}) data is used to illustrate the methodologies in a real application.

KEYWORDS. Robust estimation; PAR models; Outliers; PM_{10} pollutant.

6.1 Introduction

In general, classic time series analysis relies on stationarity assumptions. We refer to Brockwell & Davis (2006), Priestley (1981) as classical references on analysis of stationary time series. Despite the broad use of stationary tools, in some cases, this requirement is over restrictive. Examples of non-stationary phenomena are: unit roots; deterministic trends; heteroskedasticity; among others. The appropriate statistical approach will depend on the particular non-stationary phenomenon at hand.

Among these phenomena, Periodic Correlation (PC) or cyclostationarity, introduced by the seminal paper of Gladyshev (1961) deserves special attention due to the fact that it is not revealed by usual stationary techniques, which may lead to misspecification (TIAO; GRUPE, 1980). Due to this problem, special methods to identify the presence of PC have been proposed. See Hurd & Gerr (1991) or Bloomfield, Hurd & Lund (1994), for example. Since the introduction of Periodic correlation model in the literature of time series, this process has been applied in different areas of application: Gardner & Franks (1975) investigate cyclostationarity in electrical engineering; Lund et al. (1995)

have found PC in climatological time series; Noakes, McLeod & Hipel (1985) have found PC in time series of monthly river flows.

One of the most used models for PC time series is the Periodic Autoregressive (PAR) model, which generalizes the well-known Autoregressive model of Box & Jenkins (1970), where the coefficients and the orders are allowed to vary periodically in time. Estimation methods for the PAR model parameters have been studied by many authors. See Basawa & Lund (2001), Sarnaglia, Reisen & Bondon (2015) or Anderson & Meerschaert (2005), for example.

Although the PAR model has been applied in several fields, to the best of our knowledge, it still relatively unexplored in Air Pollution, especially in the context of robust problems. Among the air pollutants, the Particulate Matter with diameter smaller than $10\mu m$ (PM_{10}) is recognized for its effects on human health, see Reisen et al. (2014) and Souza et al. (2017) and references therein. This motivates using PAR models to fit PM_{10} concentrations, which is one of the focuses of this paper. It is valuable to note that, usually, Air Pollution data have asymmetric distribution and large peaks of concentrations. Therefore, classical estimators, such as the ones mentioned in the previous paragraph, are dramatically affected by the presence these observations, which suggests to consider robust estimators for PAR parameters, such as the ones proposed by Shao (2008) and Sarnaglia, Reisen & Lévy-Leduc (2010), Sarnaglia et al. (2016). However, to the best of our knowledge, there are no empirical studies investigating the behaviour of such estimators under asymmetric errors and observations which can be identified as outliers. Hence, this paper aims to investigate their finite sample properties under asymmetric errors and atypical observations through a Monte Carlo study.

A stochastic process $\{Y_t; t \in \tau\}$ is a family of random variables indexed by the set τ . If the index set is $\tau = \mathbb{Z}$ the integers set, the stochastic process is observed on equally spaced time intervals. In the following, we omit the index set \mathbb{Z} and use the short hand notation $\{Y_t\}$.

Let $\mu_t = E(Y_t)$ and $\gamma_t(h) = \text{Cov}(Y_t, Y_{t-h})$ denote the mean and Autocovariance (ACV) functions of $\{Y_t\}$, respectively. The Autocorrelation function (ACF) is defined as

$$\rho_t(h) = \frac{\gamma_t(h)}{[\gamma_t(0)\gamma_{t-h}(0)]^{\frac{1}{2}}}.$$

One says that $\{Y_t\}$ has Periodic Correlation (PC) or that it is a Periodically Stationary (PS) process if

$$\mu_{t+s} = \mu_t \quad \text{and} \quad \gamma_{t+s}(h) = \gamma_t(h), \quad t, h \in \mathbb{Z}.$$

That is, $\{Y_t\}$ is non-stationary, since the covariance structure varies under time translations. The term \mathcal{S} is called period length. We will also use $PS_{\mathcal{S}}$ to make explicit the period length of the PS process. In order to emphasize the period we are considering, we may rewrite $t = r\mathcal{S} + \nu$, $r \in \mathbb{Z}$, $\nu = 1, \dots, \mathcal{S}$, such that $\mu_{r\mathcal{S}+\nu} = \mu_{\nu}$ and $\gamma_{r\mathcal{S}+\nu}(h) = \gamma_{\nu}(h)$.

In this case, one can show that $\rho_{r\mathcal{S}+\nu}(h) = \rho_{\nu}(h)$. In addition, the Partial Autocorrelation function (PACF) defined as

$$\alpha_t(h) = \text{Corr}(Y_t, Y_{t-h} | Y_{t-1}, \dots, Y_{t-h+1})$$

is also periodic in time, that is, $\alpha_{r\mathcal{S}+\nu}(h) = \alpha_{\nu}(h)$.

The mean, ACF and PACF functions play a fundamental role on some steps of the analysis of PC time series, such as model identification. However, in practice, these functions are unknown and need to be estimated. In some applications, the period \mathcal{S} is naturally established, for example, daily data may have a weekly periodic behavior, or hourly data may present a 24-hour period, and so on. In other situations, \mathcal{S} is also an unknown parameter and needs to be inferred from the data. In this case, we can use the methods proposed by Hurd & Gerr (1991) or Bloomfield, Hurd & Lund (1994), for example, which are designed to test for the presence of PC and to estimate \mathcal{S} .

After the choice of an appropriate \mathcal{S} , estimation of the mean and the ACF functions may be carried out through period statistics. More specifically, let the sample size $N = n\mathcal{S}$ (for simplicity), that is, there are n full cycles available to perform inference. Let $Y_1, \dots, Y_{n\mathcal{S}}$ be the time series generated by a $PS_{\mathcal{S}}$ process $\{Y_t\}$. For lags $0 \leq h < [(n-1)\mathcal{S} + \nu]$, the sample (periodic) ACV is defined as

$$\hat{\gamma}_{\nu}(h) = \sum_{r=r^*}^{n-1} (Y_{r\mathcal{S}+\nu} - \bar{Y}_{\nu})(Y_{r\mathcal{S}+\nu-h} - \bar{Y}_{\nu-h}), \quad (6.1)$$

where $\bar{Y}_m = \frac{1}{n} \sum_{r=0}^{n-1} Y_{r\mathcal{S}+m}$, $m = 1, \dots, \mathcal{S}$, are the sample means, which, for $m < 1$ or $m > \mathcal{S}$, must be interpreted periodically in $m \pmod{\mathcal{S}}$, and r^* is the smallest integer such that $r^*\mathcal{S} + \nu > h$. For $-[(n-1)\mathcal{S} + \nu] < h < 0$, we access the sample ACV by using the conventions $\hat{\gamma}_{\nu}(h) = \hat{\gamma}_{\nu-h}(-h)$ and $\hat{\gamma}_{\nu+k\mathcal{S}}(h) = \hat{\gamma}_{\nu}(h)$, $k \in \mathbb{Z}$. For lags $0 \leq h < [(n-1)\mathcal{S} + \nu]$, the sample ACF is obtained from the sample ACV by

$$\hat{\rho}_{\nu}(h) = \frac{\hat{\gamma}_{\nu}(h)}{[\hat{\gamma}_{\nu}(0)\hat{\gamma}_{\nu-h}(0)]^{\frac{1}{2}}}.$$

Similarly to $\hat{\gamma}_{\nu}(h)$, for $-[(n-1)\mathcal{S} + \nu] < h < 0$, $\hat{\rho}_{\nu}(h)$ is accessed from the conventions $\hat{\rho}_{\nu}(h) = \hat{\rho}_{\nu-h}(-h)$ and $\hat{\rho}_{\nu+k\mathcal{S}}(h) = \hat{\rho}_{\nu}(h)$, $k \in \mathbb{Z}$. On the other hand, the sample PACF can be obtained through the methods in Sakai (1982) and Shao & Lund (2004), for instance.

The rest of the paper is organized as follows: Section 6.2 introduces the well-known PAR model and describes three estimation methods; Section 6.3 presents and discusses the results of the Monte Carlo experiment; Section 6.4 illustrates the use of the estimation methodologies with an application to fit Particulate Matter concentrations through a regression model with PAR errors; and Section 6.5 concludes the paper.

6.2 The PAR Model and its Estimation Methods

6.2.1 PAR Model

The Periodic Autoregressive (PAR) model is a generalization of the well-known Autoregressive (AR) process. It is one of the most used models for fitting PS_s time series. One can say that the zero-mean PS_s process $\{Y_{rS+\nu}\}$ follows a PAR model, if it satisfies the difference equation

$$Y_{rS+\nu} - \sum_{i=1}^{p_\nu} \phi_i(\nu) Y_{rS+\nu-i} = \sigma_\nu \epsilon_{rS+\nu}, \quad (6.2)$$

where $\{\epsilon_t\}$ is a white noise sequence with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = 1$, and respectively $\phi_\nu = (\phi_1(\nu), \dots, \phi_{p_\nu}(\nu))'$, p_ν and σ_ν^2 are the AR coefficient vector, the AR order and the white noise variance at the period ν .

Observe that, there is no loss of generality in considering constant AR orders, say $p_1 = \dots = p_s = p$. In order to see that, take $p = \max_\nu p_\nu$ and restrict the additional AR coefficients to zero, that is, $\phi_\nu(i) = 0$, $i > p_\nu$, $\nu = 1, \dots, S$. Therefore, from this point on, we shall consider the AR orders constant over the S periods.

The well-known causal representation can be extended to the periodic scenario. One says that the zero-mean PS_s process $\{Y_{rS+\nu}\}$ is (periodically) causal if it can be written as

$$Y_{rS+\nu} = \sum_{j \geq 0} \psi_j(\nu) \epsilon_{rS+\nu-j},$$

where $\max_\nu [\sum_{j \geq 0} |\psi_j(\nu)|] < \infty$. Particular conditions to ensure causality of a PAR model can be derived from its vector AR representation. See Sarnaglia, Reisen & Lévy-Leduc (2010), for example. For a PAR(1) model, the causality condition is reduced to

$$\lambda = \left| \prod_{\nu=1}^S \phi_1(\nu) \right| < 1. \quad (6.3)$$

The invertible representation of a PS process can be defined in similar lines as

$$\sum_{j \geq 0} \pi_j(\nu) Y_{rS+\nu-j} = \epsilon_{rS+\nu},$$

where $\max_\nu [\sum_{j \geq 0} |\pi_j(\nu)|] < \infty$. It is straightforward to see that, as in the stationary AR case, the PAR model is always invertible.

Other aspects of PAR models can be found in Sarnaglia, Reisen & Lévy-Leduc (2010) and references therein.

6.2.2 Estimation Methods

We now describe three methodologies to perform the estimation of PAR models. The first, is a generalization of the Yule-Walker estimator of stationary AR models. The other

ones are robust estimators recently proposed in literature. Henceforth, in order to keep the notation clean, we shall group all the autoregressive coefficients in Equation 6.2 inside the vector $\boldsymbol{\phi} = (\boldsymbol{\phi}'_1, \dots, \boldsymbol{\phi}'_S)' = (\phi_1(1), \dots, \phi_p(1), \dots, \phi_1(S), \dots, \phi_p(S))'$, where, as previously, for $\nu = 1, \dots, S$, $\boldsymbol{\phi}_\nu = (\phi_1(\nu), \dots, \phi_p(\nu))'$, denotes the vector of AR coefficients of the period ν .

6.2.2.1 Yule-Walker Estimator

A simple estimator can be developed through the method of moments. For a fixed $\nu = 1, \dots, S$, by multiplying $Y_{rS+\nu-h}$ and taking expectations in both sides of (6.2), one obtains the following generalization of the Yule-Walker equations to the PC scenario:

$$\sum_{i=1}^p \phi_i(\nu) \gamma_{\nu-i}(h-i) = \gamma_\nu(h), \quad h = 1, \dots, p. \quad (6.4)$$

The Yule-Walker Estimator (YWE) of the PAR coefficients is defined as follows. For each $\nu = 1, \dots, S$, replace, in (6.4), the theoretical ACV, $\gamma_\nu(h)$, by its sample counterpart (Equation 6.1), $\hat{\gamma}_\nu(h)$, and solve the resulting linear equations system

$$\sum_{i=1}^p \phi_i(\nu) \hat{\gamma}_{\nu-i}(h-i) = \hat{\gamma}_\nu(h), \quad k = 1, \dots, p,$$

for $\phi_1(\nu), \dots, \phi_p(\nu)$. The solution $\hat{\boldsymbol{\phi}}_\nu = (\hat{\phi}_1(\nu), \dots, \hat{\phi}_p(\nu))'$ is the YWE of the AR coefficients of the period ν . Thus, the YWE of $\boldsymbol{\phi}$ is defined as $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}'_1, \dots, \hat{\boldsymbol{\phi}}'_S)'$. By taking $h = 0$, the white noise variance σ_ν^2 can also be estimated by the same argument as in Equation 6.4.

Note that the YWE results of solving S successive systems of linear equations. Therefore, provided that each matrix of the systems is invertible, the YWE can be analytically obtained, which ensures small computational effort.

Asymptotics results for YWE can be derived under the following:

Assumption 1. $\{Y_t\}$ is a zero-mean causal PAR process and $E(Y_t^4) < \infty$, $t \in \mathbb{Z}$.

The following result has been obtained by Sarnaglia, Reisen & Lévy-Leduc (2010):

Remark 1. Under Assumption 1, the YWE, $\hat{\boldsymbol{\phi}}$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \rightsquigarrow \mathcal{N}_{pS}(0, G), \quad n \rightarrow \infty,$$

where the \rightsquigarrow symbol denotes convergence in distribution, and the covariance matrix G can be found on Proposition 1 of Sarnaglia, Reisen & Lévy-Leduc (2010).

Other asymptotically equivalent estimators of PAR coefficients are the Least Squares Estimator (LSE) (BASAWA; LUND, 2001) and the Maximum Likelihood Estimator

(MLE) (LUND; BASAWA, 1999). Due to that equivalence, these estimators will not be considered in this paper.

Many simulation results (SHAO, 2008; SARNAGLIA; REISEN; LÉVY-LEDUC, 2010) reveal classic estimators, such as YWE, LSE and MLE, are not resistant in the presence of atypical observations, such as Additive Outliers (AO). This problem has motivated investigation of robust approaches in literature. In Subsections 6.2.2.2 and 6.2.2.3, we describe the particular methods proposed by Sarnaglia, Reisen & Lévy-Leduc (2010) and Shao (2008), respectively.

6.2.2.2 Robust Yule-Walker Estimator

In order to prevent the harmful effects of outliers, Sarnaglia, Reisen & Lévy-Leduc (2010) have proposed a Robust Yule-Walker Estimator (RYWE). Firstly, the authors generalize the robust autocovariance function proposed in Ma & Genton (2000) to the periodic scenario as follows. Let $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$, the robust scale estimator of \mathbf{y} proposed by Rousseeuw & Croux (1993), $Q_n(\mathbf{y})$, is defined as the following k th order statistic

$$Q_n(\mathbf{y}) = d \{ |y_i - y_j| ; 1 \leq i < j \leq n \}_{(k)}, \quad (6.5)$$

where d is a constant factor to ensure Fisher-consistency and $k = \binom{c}{2} \approx 0.25 \binom{n}{2}$, where $c = \lfloor n/2 \rfloor + 1$ is roughly half of the size n of the vector \mathbf{y} . For Gaussian random variables, $d = 2.2191$. Given a PS_s time series, Y_1, \dots, Y_{n_s} , based on (6.5) and, following Ma & Genton (2000), Sarnaglia, Reisen & Lévy-Leduc (2010) define the robust sample (periodic) ACV function by

$$\tilde{\gamma}_\nu(h) = \frac{1}{4} [Q_{n-r^*+1}^2(\mathbf{u}_\nu + \mathbf{v}_\nu) - Q_{n-r^*+1}^2(\mathbf{u}_\nu - \mathbf{v}_\nu)], \quad 0 \leq h < [(n-1)s + \nu], \quad (6.6)$$

where $\mathbf{u}_\nu = (Y_{r^*s+\nu-h}, \dots, Y_{(n-1)s+\nu-h})$, $\mathbf{v}_\nu = (Y_{r^*s+\nu}, \dots, Y_{(n-1)s+\nu})$ and r^* was previously defined in Equation 6.1. Definition of $\tilde{\gamma}_\nu(h)$ for h in other lag ranges follows similarly to definition of $\hat{\gamma}_\nu(h)$. The main disadvantage of the robust ACV is that it is not semidefinite positive.

The RYWE is defined similarly to the YWE in the previous section. For each $\nu = 1, \dots, s$, replace, in (6.4), the theoretical ACV, $\gamma_\nu(h)$, by its sample robust estimator in Equation 6.6, $\tilde{\gamma}_\nu(h)$, and solve the resulting linear equations system

$$\sum_{i=1}^p \phi_i(\nu) \tilde{\gamma}_{\nu-i}(h-i) = \tilde{\gamma}_\nu(h), \quad k = 1, \dots, p, \quad (6.7)$$

for $\phi_1(\nu), \dots, \phi_p(\nu)$. Note that the above systems do not depend on the Fisher-consistency constant d in (6.5). The solution of the system in (6.7), $\tilde{\boldsymbol{\phi}}_\nu = (\tilde{\phi}_1(\nu), \dots, \tilde{\phi}_p(\nu))'$, is the RYWE of the AR coefficients of the period ν . The RYWE is defined as $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\phi}}_1', \dots, \tilde{\boldsymbol{\phi}}_s')'$. By taking $h = 0$, the white noise variance σ_ν^2 can also be robustly estimated by the same argument as in Equation 6.4.

Similarly as for YWE, the RYWE results of solving \mathcal{S} successive systems of linear equations. Therefore, provided that each matrix of the systems is invertible, there also exists an analytic closed expression to the RYWE, which ensures little computational effort to provide the estimates. However, the computation of $\tilde{\phi}$ might demand more time compared to the evaluation of $\hat{\phi}$, sorting operations are required to evaluate the order statistic $Q_n(\cdot)$.

In order to access the rate of convergence of the RYWE, the following assumption is imposed:

Assumption 2. For any $\nu = 1, \dots, \mathcal{S}$, $\{Y_{r\mathcal{S}+\nu}; r \in \mathbb{Z}\}$ is a mean-zero Gaussian process with strong mixing coefficients α_n satisfying: $\alpha_n \leq Cn^{-a}$, for some $a > 1$ and $C \geq 1$.

The following result is due to Sarnaglia, Reisen & Lévy-Leduc (2010):

Remark 2. Under Assumptions 1 and 2, the RYWE $\tilde{\phi}$ satisfies

$$\tilde{\phi}_i(\nu) - \phi_i(\nu) = O_p(n^{-1/2}),$$

for $i = 1, \dots, p$, $\nu = 1, \dots, \mathcal{S}$.

6.2.2.3 Robust Least Squares Estimator

In an effort to improve robustness, Shao (2008) proposes an alternative to the conditional Least Squares Estimator (LSE). The LSE of ϕ is defined as

$$\underset{\phi \in \Phi}{\operatorname{argmin}}[L(\phi)],$$

where Φ is the appropriate parameter space and the objective function $L(\phi)$ is given by

$$L(\phi) = \sum_{r=0}^{n-1} \sum_{\nu=1}^{\mathcal{S}} \frac{1}{\sigma_\nu^2} \left(Y_{r\mathcal{S}+\nu} - \sum_{i=1}^p \phi_i(\nu) Y_{r\mathcal{S}+\nu-i} \right)^2, \quad (6.8)$$

with $Y_t = 0$, $t \leq 0$, by convention. Corresponding to L in (6.8), one can obtain the $p\mathcal{S}$ -dimensional estimating equations

$$S_n(\phi_\nu) = \frac{1}{\sigma_\nu} \sum_{r=0}^{n-1} \epsilon_{r\mathcal{S}+\nu} Y_{r\mathcal{S}+\nu-i} = 0, \quad 1 \leq i \leq p, \quad 1 \leq \nu \leq \mathcal{S}, \quad (6.9)$$

where the error terms are given by $\epsilon_{r\mathcal{S}+\nu} = \epsilon_{r\mathcal{S}+\nu}(\phi_\nu) = (Y_{r\mathcal{S}+\nu} - \sum_{i=1}^p \phi_i(\nu) Y_{r\mathcal{S}+\nu-i})/\sigma_\nu$, $0 \leq r < n$, $1 \leq \nu \leq \mathcal{S}$. Asymptotics for the LSE have been studied by Basawa & Lund (2001).

Aiming to improve robustness of the LSE, Shao (2008) proposes to solve a modification of the estimating equations (6.9) by replacing, respectively, $\epsilon_{r\mathcal{S}+\nu}$ and $Y_{r\mathcal{S}+\nu}$ by their robust versions $\check{\epsilon}_{r\mathcal{S}+\nu}$ and $\check{Y}_{r\mathcal{S}+\nu}$ defined as

$$\check{\epsilon}_{r\mathcal{S}+\nu} = \psi(\epsilon_{r\mathcal{S}+\nu}) \quad (6.10)$$

and

$$\check{Y}_{rS+\nu} = \begin{cases} Y_{rS+\nu}, & \text{if } \check{\epsilon}_{rS+\nu} = \epsilon_{rS+\nu}, \\ \sum_{i=1}^p \phi_i(\nu) \check{Y}_{rS+\nu-i} + \sigma_\nu \check{\epsilon}_{rS+\nu}, & \text{if } \check{\epsilon}_{rS+\nu} \neq \epsilon_{rS+\nu}. \end{cases} \quad (6.11)$$

Shao (2008) has considered $\psi(\cdot)$ as the Huber type function, defined by

$$\psi(x) = \psi_c(x) = \begin{cases} x, & \text{if } |x| \leq c, \\ c \operatorname{sign}(x), & \text{if } |x| > c. \end{cases}$$

Under the additional assumption that the marginal density function $f_\epsilon(\cdot)$ of the error term ϵ_t is symmetric about zero, it can be shown that $E(\check{\epsilon}_t) = E(\psi(\epsilon_t)) = 0$. The Robust LSE (RLSE) is the solution of the robustified estimating equations

$$\check{S}_n(\phi_\nu) = \frac{1}{\sigma_\nu} \sum_{r=0}^{n-1} \psi \left(\frac{Y_{rS+\nu} - \sum_{i=1}^p \phi_i(\nu) Y_{rS+\nu-i}}{\sigma_\nu} \right) \check{Y}_{rS+\nu-i} = 0, \quad 1 \leq i \leq p, \quad 1 \leq \nu \leq S. \quad (6.12)$$

Although Shao (2008) uses ψ as the Huber function, he has emphasized that any odd bounded and differentiable function can be a candidate for the $\psi(\cdot)$ function. A Huber type function has the advantage of being monotone, which ensures the existence and uniqueness of solution to Equation 6.12. However, the drawback of the so-called redescending, such as Bisquare, Hampel, Generalized Gauss-Weight, among others, is the existence of many zeroes, which may lead to non-optimal solutions.

In practice, the estimates are obtained in the following iterative procedure starting with an appropriate initial guess for the RLSE. Suppose $\check{\phi}^{(l)}$ represents the vector of estimates at the l th iteration. Then, at the $(l+1)$ th iteration, calculate the residuals

$$e_{rS+\nu}^{(l)} = Y_{rS+\nu} - \sum_{i=1}^p \check{\phi}_i^{(l)}(\nu) Y_{rS+\nu-i}, \quad 1 \leq \nu \leq S,$$

where $Y_t = 0$, $t \leq 0$, estimate the white noise standard deviation at the period ν , σ_ν , by

$$\check{\sigma}_\nu^{(l)} = \operatorname{Median} \left(|e_\nu^{(l)}|, |e_{S+\nu}^{(l)}|, \dots, |e_{(n-1)S+\nu}^{(l)}| \right), \quad 1 \leq \nu \leq S,$$

calculate the robust version of $e_{rS+\nu}^{(l)}$ through (6.10), $\check{e}_{rS+\nu}^{(l)} = \psi(e_{rS+\nu}^{(l)})$, obtain $\check{Y}_{rS+\nu}^{(l)}$ by (6.11) with $\check{\epsilon}_{rS+\nu}$ substituted for the robust residual $\check{e}_{rS+\nu}^{(l)}$ and σ_ν replaced with $\check{\sigma}_\nu^{(l)}$, and evaluate the solution $\check{\phi}^{(l+1)}$ of the robustified estimating equations in (6.12) replacing $\check{Y}_{rS+\nu}$ with $\check{Y}_{rS+\nu}^{(l)}$ and σ_ν with $\check{\sigma}_\nu^{(l)}$. Stop the procedure according to some convergence criterion.

In order to verify asymptotics of $\check{\phi}$, the following assumption is required:

Assumption 3. The marginal density function $f_\epsilon(\cdot)$ of the error $\epsilon_{rS+\nu}$ in Equation 6.2 is symmetric about the origin.

The following result has been derived by Shao (2008):

Remark 3. Under Assumptions 1 and 3, the RLSE $\check{\phi}$ satisfies

$$\sqrt{n}(\check{\phi} - \phi) \rightsquigarrow \mathcal{N}_{pS}(0, A), \quad n \rightarrow \infty,$$

where the covariance matrix A is given in Equation 14 of Shao (2008).

6.3 Simulation

6.3.1 Specification of the Monte Carlo Experiment

In order to investigate the impact of atypical observations on the estimation methods, series were generated with and without additive outliers. The AO contaminated PS process is defined as follows. $\{X_{rS+\nu}\}$ is a periodically stationary process contaminated with additive outliers if it satisfies

$$X_{rS+\nu} = Y_{rS+\nu} + \sum_{i=1}^{\mathcal{O}} \omega_i V_{i,rS+\nu} \quad (6.13)$$

where $\{Y_{rS+\nu}\}$ is a periodically stationary process, $\{V_{i,t}\}$ is a sequence of independent random variables with $P(V_{i,t} = -1) = P(V_{i,t} = 1) = \xi_i/2$ and $P(V_{i,t} = 0) = 1 - \xi_i$, $0 \leq \xi_i < 1$, Y_t and V_s are independent processes for all t, s and \mathcal{O} is the number of different outliers and ω_i is the magnitude of the i th type of outlier.

The simulation study is divided in these two cases; uncontaminated and contaminated series. The contaminated series was generated from model in (6.13) with the following specifications: $\mathcal{O} = 1$; $\omega_1 = \omega = 7$; and $\xi_1 = \xi = 0.01$. The effect of the normality departure in the white noise sequence was also studied by generating the random variables ϵ_t such that $\epsilon \sim \mathcal{N}(0, 1)$ and $\sqrt{2}\epsilon_t + 1 \sim \chi_{(1)}^2$. Other cases of outliers were simulated as well as other non-normal distributions. However, since the conclusions are similar to the ones discussed here the results are available upon request. For each of these scenarios, 1000 replicates of $\{Y_t\}$ were generated following the PAR models with $S = 4$ and coefficients given in Table 3. Model 2 is closer to the non-causality region than Model 1. The sample sizes were taken as $N = 400$ and 1600 ($n = 100$ and 400 cycles, respectively). The initial values for RLSE was taken as the true parameter vector.

Table 3 – Parameters of PAR(1) models used in the simulation.

Parameter	Model 1	Model 2
$\phi_1(1)$	0.9	1.5
$\phi_1(2)$	0.8	0.8
$\phi_1(3)$	0.7	1.2
$\phi_1(4)$	0.6	0.5
λ	0.3024	0.7200

6.3.2 Results

The performance of the estimators is accessed through the empirical Bias and Root Mean Squared Error (RMSE). Tables 4 and 5 display the Bias and RMSE for Models 1 and 2, respectively. For illustration purpose, the empirical distributions of $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$, $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ and $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ for Model 1, with $n = 400$, under the uncontaminated and contaminated Gaussian scenarios are presented in Figures 3 and 4, respectively. The same plots for the uncontaminated case with asymmetric errors are displayed in Figure 5.

Table 4 – Bias and RMSE for Model 1 and outliers with probability $\xi = 0.01$.

ω	ϵ_t	n	$\phi_1(\nu)$	YWE		RYWE		RLSE	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
0	$\mathcal{N}(0, 1)$	100	0.9	-0.007	0.077	-0.003	0.103	0.003	0.079
			0.8	-0.002	0.065	0.004	0.084	-0.002	0.065
			0.7	0.000	0.063	-0.001	0.083	-0.001	0.063
			0.6	-0.005	0.066	-0.003	0.083	-0.005	0.067
		400	0.9	-0.001	0.037	-0.001	0.047	0.002	0.038
			0.8	-0.001	0.031	0.000	0.038	-0.001	0.031
			0.7	-0.001	0.032	0.001	0.038	-0.001	0.032
			0.6	0.000	0.032	0.000	0.039	0.000	0.033
	$\frac{\chi^2_{(1)} - 1}{\sqrt{2}}$	100	0.9	-0.006	0.076	0.178	0.209	0.006	0.063
			0.8	-0.007	0.065	0.117	0.147	-0.003	0.055
			0.7	-0.004	0.065	0.088	0.119	-0.003	0.052
			0.6	-0.005	0.069	0.077	0.108	-0.004	0.056
		400	0.9	-0.001	0.037	0.179	0.185	0.002	0.030
			0.8	0.000	0.033	0.115	0.122	0.000	0.026
			0.7	-0.001	0.033	0.089	0.096	-0.001	0.026
			0.6	-0.001	0.034	0.085	0.093	0.000	0.028
7	$\mathcal{N}(0, 1)$	100	0.9	-0.181	0.247	0.014	0.120	-0.031	0.095
			0.8	-0.118	0.176	0.012	0.096	-0.027	0.076
			0.7	-0.105	0.157	0.015	0.091	-0.024	0.077
			0.6	-0.097	0.151	0.012	0.091	-0.027	0.081
		400	0.9	-0.183	0.203	0.017	0.055	-0.027	0.050
			0.8	-0.129	0.144	0.012	0.046	-0.021	0.041
			0.7	-0.108	0.124	0.013	0.044	-0.019	0.041
			0.6	-0.103	0.119	0.014	0.043	-0.022	0.043
	$\frac{\chi^2_{(1)} - 1}{\sqrt{2}}$	100	0.9	-0.172	0.243	0.213	0.251	-0.019	0.076
			0.8	-0.126	0.180	0.142	0.175	-0.021	0.063
			0.7	-0.105	0.158	0.112	0.144	-0.018	0.061
			0.6	-0.096	0.151	0.103	0.134	-0.023	0.067
		400	0.9	-0.182	0.202	0.211	0.219	-0.021	0.041
			0.8	-0.129	0.145	0.142	0.149	-0.017	0.033
			0.7	-0.112	0.127	0.111	0.119	-0.017	0.033
			0.6	-0.106	0.121	0.106	0.114	-0.019	0.037

Table 5 – Bias and RMSE for Model 2 and outliers with probability $\xi = 0.01$.

ω	ϵ_t	n	$\phi_1(\nu)$	YWE		RYWE		RLSE	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
0	$\mathcal{N}(0, 1)$	100	1.5	-0.009	0.055	-0.009	0.100	0.000	0.055
			0.8	-0.004	0.033	-0.005	0.050	-0.004	0.034
			1.2	-0.006	0.040	-0.008	0.066	-0.006	0.040
			0.5	-0.008	0.033	-0.008	0.043	-0.008	0.033
		400	1.5	-0.003	0.026	-0.003	0.037	-0.001	0.026
			0.8	-0.001	0.016	-0.001	0.021	-0.001	0.016
			1.2	-0.002	0.019	-0.002	0.027	-0.002	0.020
			0.5	-0.002	0.015	-0.003	0.018	-0.003	0.015
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	1.5	-0.009	0.055	0.025	0.098	0.001	0.043
			0.8	-0.007	0.036	0.011	0.048	-0.005	0.028
			1.2	-0.007	0.040	0.019	0.069	-0.005	0.033
			0.5	-0.009	0.033	-0.004	0.039	-0.006	0.027
		400	1.5	-0.002	0.026	0.032	0.050	0.000	0.021
			0.8	-0.001	0.016	0.016	0.026	-0.001	0.013
			1.2	-0.002	0.019	0.023	0.035	-0.002	0.015
			0.5	-0.003	0.015	0.003	0.017	-0.002	0.012
7	$\mathcal{N}(0, 1)$	100	1.5	-0.174	0.246	-0.023	0.122	-0.014	0.061
			0.8	-0.047	0.076	-0.011	0.058	-0.014	0.040
			1.2	-0.079	0.125	-0.013	0.079	-0.013	0.048
			0.5	-0.029	0.056	-0.010	0.046	-0.013	0.038
		400	1.5	-0.167	0.189	-0.020	0.052	-0.013	0.032
			0.8	-0.042	0.050	-0.007	0.025	-0.007	0.019
			1.2	-0.080	0.093	-0.015	0.038	-0.008	0.022
			0.5	-0.023	0.032	-0.006	0.021	-0.007	0.018
	$\frac{\chi^2_{(1)}-1}{\sqrt{2}}$	100	1.5	-0.180	0.253	0.022	0.120	-0.013	0.054
			0.8	-0.049	0.081	0.011	0.057	-0.008	0.032
			1.2	-0.082	0.126	0.016	0.083	-0.009	0.036
			0.5	-0.029	0.053	-0.003	0.041	-0.010	0.028
		400	1.5	-0.172	0.193	0.022	0.054	-0.010	0.025
			0.8	-0.041	0.050	0.015	0.028	-0.005	0.015
			1.2	-0.083	0.097	0.012	0.039	-0.007	0.018
			0.5	-0.024	0.033	0.003	0.020	-0.005	0.014

For Model 1 (Table 4), under the uncontaminated Gaussian case, the reduction of Bias and RMSE when increasing the sample size suggests that all estimators are consistent. We observe that, in this scenario, YWE and RLSE present better results. The findings for the asymmetric uncontaminated case show that the RYWE is extremely affected by this phenomenon, presenting a persistent Bias which does not seem to vanish by increasing the sample size. The results for heavy tailed errors are similar, which show that the RYWE is very sensitive to departures from normality. This problem indicates that the normality requirement in Assumption 2 is crucial to ensure asymptotic properties of the RYWE. The YWE and RLSE do not seem to be affected by non-Gaussian errors (both asymmetric and

heavy tailed). This gives empirical evidence that the technical requirement of symmetric errors (Assumption 3) may be over restrictive to ensure asymptotic normality of the RLSE. As expected, atypical observations increase the Bias and RMSE of the YWE. Under normally distributed errors, both RYWE and RLSE show robustness with Bias and RMSE almost unchanged with the presence of outliers. For asymmetric errors, only the RLSE have good performance.

The conclusions for Model 2 (Table 5) are almost the same. It is worth noting that in this stronger dependence scenario, there is a overall reduction of Bias and RMSE. Another remarkable fact is that, in this case, the RYWE does not seem to be affected by asymmetric errors in both uncontaminated and contaminated scenarios.

From Figure 3, we can see that the empirical distributions are virtually the same. However, the scale of the RYWE distribution is slightly greater than of those of the YWE and RLSE. Figure 4 shows the robustness to outliers of RYWE and RLSE under Gaussian errors, while the YWE distribution is shifted to the left due to the well-known memory loss property. Its scale is also increased as a result of the contamination. Figure 5 illustrates the prominent shift to the right of the RYWE distribution caused by the asymmetry of the errors.

6.4 Application

In order to illustrate the methodologies described in this paper, we consider an application to air pollution in the Region of Greater Vitória (RGV). RGV is formed by seven cities: Cariacica, Fundão, Guarapari, Serra, Viana, Vila Velha and Vitória. RGV is located at the Espírito Santo (ES) state, Brazil, and represents 3.2% and 43% of the ES area and population, respectively. RGV is responsible for most (from 55% to 65%) of the polluting activities in the state. In ES, the State Institute of Environment and Water Resources (IEMA) is the entity responsible for planning, coordinating, executing, overseeing and controlling environmental activities, among other assignments. Given the importance of RGV, IEMA has installed the Greater Vitória Automatic Air Quality Monitoring Network (RAMQAr), which is composed by nine monitoring stations placed in strategic locations and accounts for the measuring of several atmospheric pollutants and meteorological variables in RGV. More specifically, we consider the data set of Particulate Matter with diameter smaller than $10\mu m$ (PM_{10}) collected by the RAMQAr station located at Enseada do Suá. RAMQAr collects the atmospheric pollutant PM_{10} continuously and measures its concentration hourly, in $\mu g/m^3$.

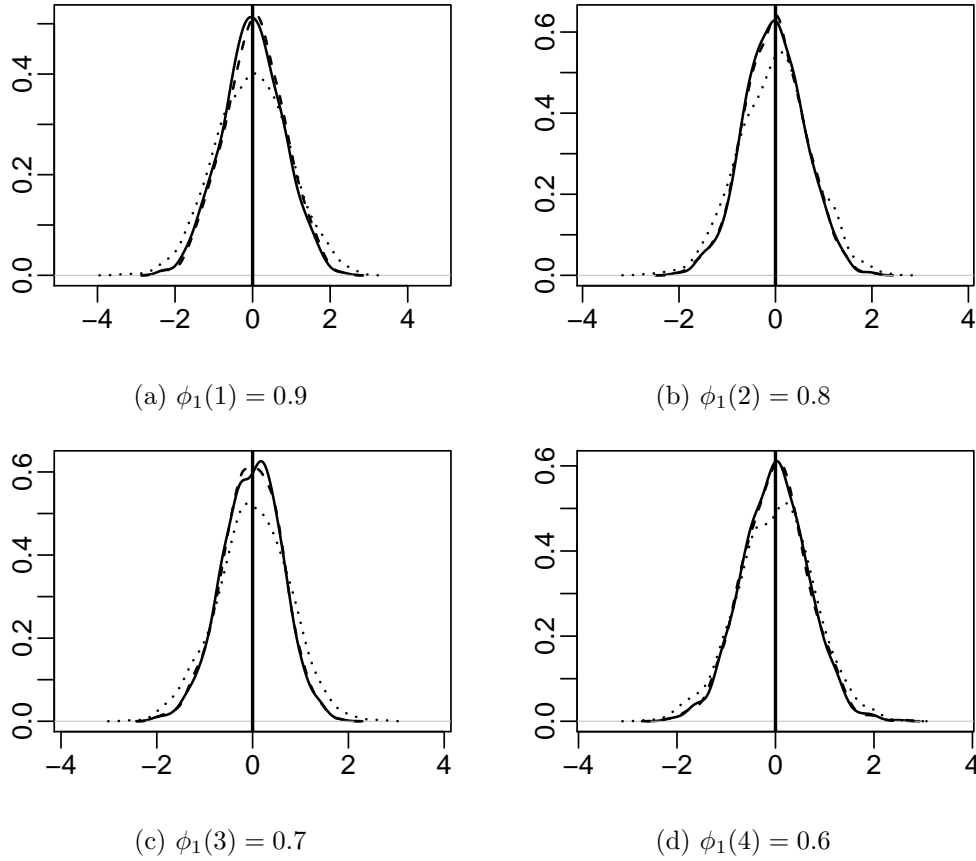


Figure 3 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 0$, $n = 400$ and normal errors.

6.4.1 Exploratory Analysis

The missing data in the PM_{10} hourly time series were imputed through the `mice` package of software R, jointly with other pollutant and meteorologic variables. We consider the data from January 1st, 2014 to December 29th, 2015. We did not take the full year of 2015 in order to keep the sample size multiple of the natural choice to period length, $\mathcal{S} = 7$. The hourly data were aggregated into daily averages. Due to asymmetry and evidence of time varying variance, we have taken the log of the observations. The plot of the $\log(PM_{10})$ is displayed in Figure 6. From this figure, we can see large peaks of PM_{10} concentration which can be viewed here as outliers, since their values can provoke serious damage to the statistical functions, such as the mean and the standard deviation and, therefore, may affect the correlation structure of the series, leading to misleading results. It also can be seen the presence of sinusoidal deterministic trends. Analysis of the periodogram (Figure 7) corroborates to this result and indicates that the frequency $2/N$, corresponding approximately to a yearly cycle, has a large contribution to the overall variance of the data. The high frequency peaks of the periodogram correspond to weekly periodicity and, according to the daily periodic boxplots displayed in Figure 8, they can

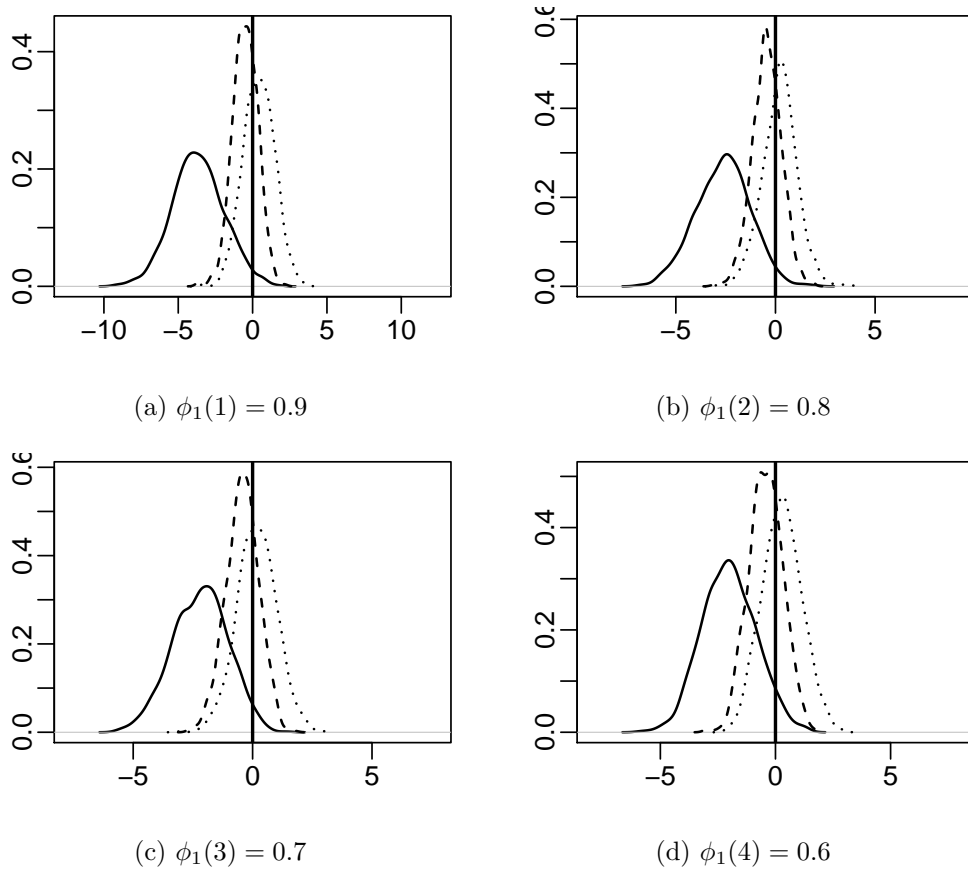


Figure 4 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 7$, $n = 400$ and normal errors.

be explained by a level decrease on the weekends. This is an expected finding due to the traffic and civil construction decreases in the region on these days.

The results above suggest that a deterministic trend must be firstly removed from $\log(PM_{10})$ before further analysis. In the next section, based on the descriptive results, we fit a linear model with the proper covariates and errors following a PAR model. The parameters of the PAR fit will be estimated through the estimation methods described earlier.

6.4.2 Linear Model with PAR Errors

Based on the results of the previous section, we define the sinusoidal covariate: $\cos_t = \cos(\frac{2\pi t}{365.25})$, $t = 1, \dots, N$; the linear term t ; and a “day of the week” factor with the levels: Week (the reference level); Saturday (represented by the dummy variable sat_t which takes value 1 for Saturdays); and Sunday (sun_t which takes value 1 for Sundays). We also considered $\sin_t = \sin(\frac{2\pi t}{365.25})$, $t = 1, \dots, N$, which turn out to be insignificant.

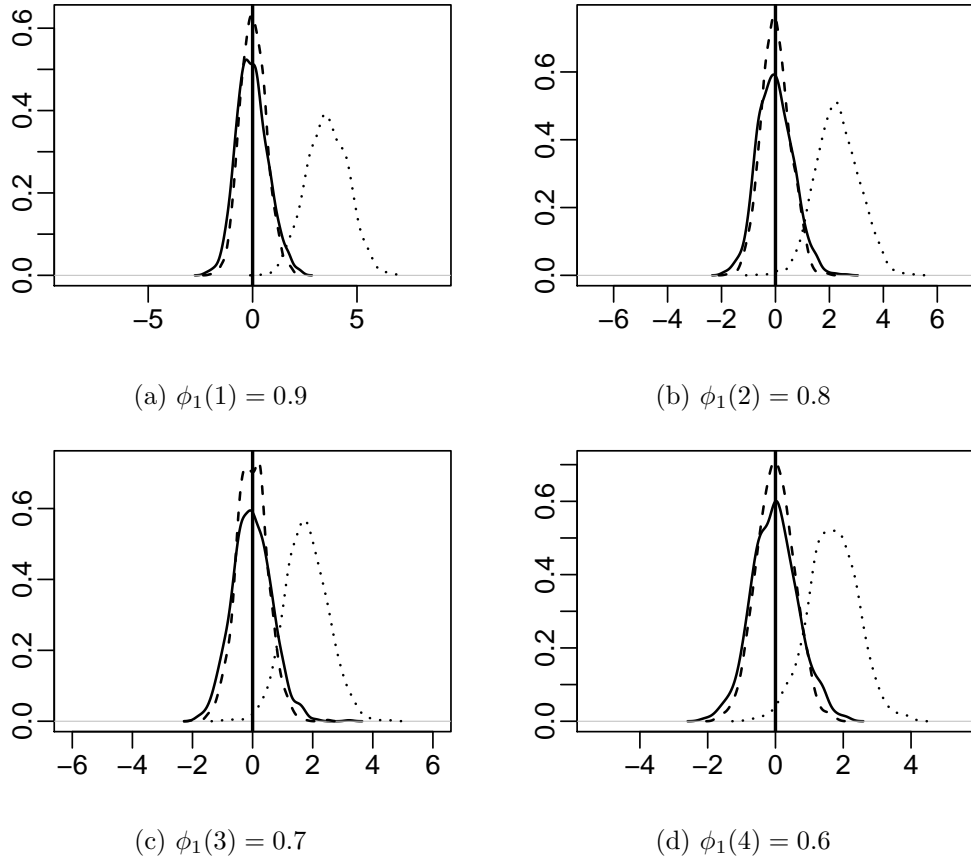


Figure 5 – Empirical Distribution of the $\sqrt{n}(\hat{\phi}_i(\nu) - \phi_i(\nu))$ (solid lines), $\sqrt{n}(\tilde{\phi}_i(\nu) - \phi_i(\nu))$ (dotted lines) and the $\sqrt{n}(\check{\phi}_i(\nu) - \phi_i(\nu))$ (dashed lines) for Model 1 with $\omega = 0$, $n = 400$ and asymmetric errors.

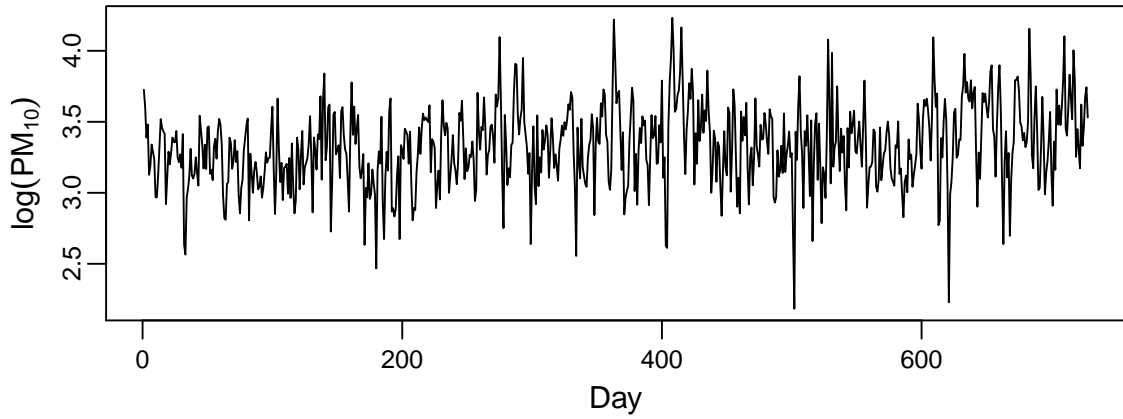


Figure 6 – Evolution of the $\log(PM_{10})$ time series.

Therefore, we shall consider the following model:

$$\log(PM_{10,t}) = \mu + \alpha_1 \text{sat}_t + \alpha_2 \text{sun}_t + \beta_1 \cos_t + \beta_2 t + Y_t; \quad (6.14)$$

$$Y_t = \sum_{i=1}^{p_t} \phi_i(t) Y_{t-i} + \sigma_t \epsilon_t, \quad (6.15)$$

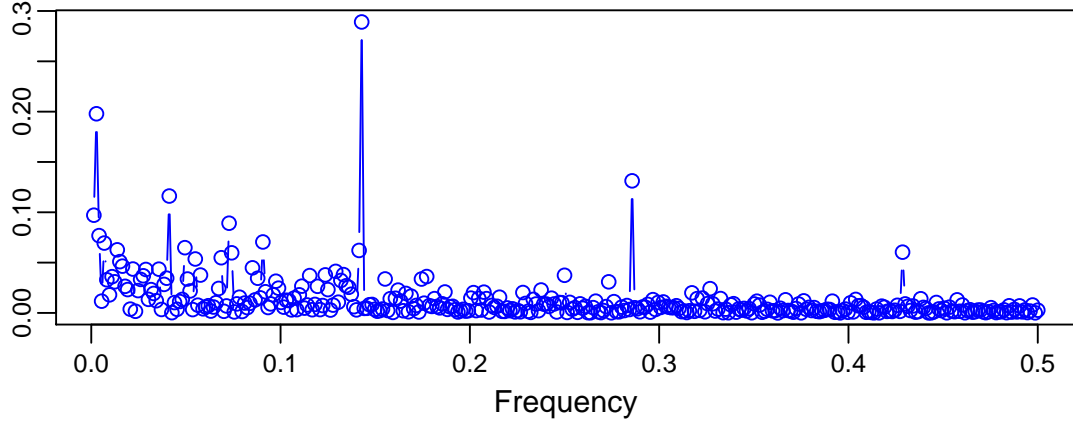


Figure 7 – Periodogram of the $\log(PM_{10})$ time series.

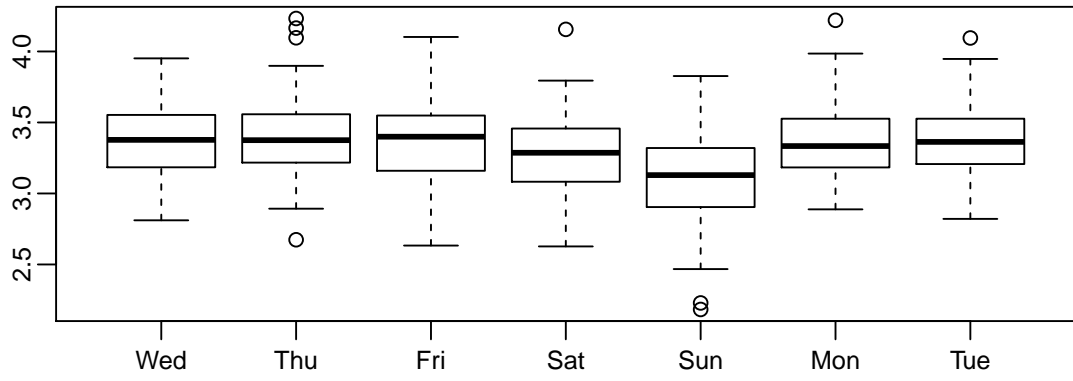


Figure 8 – Daily box-plots of the $\log(PM_{10})$ time series.

$t = 1, \dots, 728$. Due to the daily nature of the data, we will take $\mathcal{S} = 7$. The above model means that on business days the regular level of $\log(PM_{10})$ is μ , on Saturdays it suffers an increase of α_1 and on Sundays it is increased by α_2 . This level is also dependent on a long-run cyclic trend, represented by \cos_t and a linear term t .

The model in Equations 6.14 and 6.15 will be fitted in the following two steps procedure: (1) the linear model in (6.14) will be estimated through the ordinary least squares procedure; and (2) the PAR model in (6.15) will be fitted to the residuals of the linear model in step (1), where the AR orders p_1, \dots, p_s will be identified through the Schwartz Information Criterion (BIC) proposed by Schwarz (1978) and adapted to the periodic scenario by McLeod (1994).

Before the analysis, we describe the BIC procedure. The BIC is statistic to establish a compromise between quality of adjustment and overparametrization. In this case, we define the BIC as

$$\text{BIC} = \sum_{\nu=1}^{\mathcal{S}} \text{BIC}_{\nu}, \quad \text{BIC}_{\nu} = n \log(\sigma_{\nu}^{2\dagger}) + p_{\nu} \log(n), \quad \nu = 1, \dots, \mathcal{S}. \quad (6.16)$$

where $\sigma_{\nu}^{2\dagger}$ denotes an estimator of the white noise variance. The “better” model is the one which minimize the BIC. For the YWE we have used $\sigma_{\nu}^{2\dagger} = \hat{\sigma}_{\nu}^2 = \text{RSS}_{\nu}/n$, where

RSS_ν denotes the period ν Residual Sum of Squares, that is $RSS_\nu = \sum_{r=0}^{n-1} \hat{e}_{rS+\nu}^2$, wherein $\hat{e}_{rS+\nu} = Y_{rS+\nu} - \sum_{i=1}^{p_\nu} \hat{\phi}_i(\nu) Y_{rS+\nu-i}$ are the residuals of the YWE. The residuals for the RYWE and the RLSE are evaluated similarly, with $\hat{\phi}_i(\nu)$ replaced by the appropriate estimator. However, in order to improve robustness, we change the $\sigma_\nu^{2\dagger}$ estimator. For these estimators, we take

$$\sigma_\nu^{2\dagger} = \tilde{\gamma}_{\nu,e}(0), \quad (6.17)$$

where: for the RYWE, the robust ACV in 6.17, $\tilde{\gamma}_{\nu,e}(0)$, is evaluated as in Equation 6.5, with the vectors $\mathbf{u}_\nu = (\tilde{e}_{r^*S+\nu-h}, \dots, \tilde{e}_{nS+\nu-h})$ and $\mathbf{v}_\nu = (\tilde{e}_{r^*S+\nu}, \dots, \tilde{e}_{nS+\nu})$, wherein \tilde{e}_t correspond to the residuals of the RYWE fit; while, for the RLSE, the robust ACV in 6.17, $\tilde{\gamma}_{\nu,e}(0)$, is evaluated as in Equation 6.5, with the vectors $\mathbf{u}_\nu = (\check{e}_{r^*S+\nu-h}, \dots, \check{e}_{nS+\nu-h})$ and $\mathbf{v}_\nu = (\check{e}_{r^*S+\nu}, \dots, \check{e}_{nS+\nu})$, wherein \check{e}_t are the residuals resulting from the RLSE fit. We prefer to use the Q_n statistic to estimate σ_ν instead of the standard Mean Absolute Deviation (MAD), since Q_n has greater efficiency on Gaussian samples compared to the MAD (ROUSSEEUW; CROUX, 1993).

It is important to point out that, due to the need of numerical optimization, the use of RLSE on a repetitive context such as in BIC evaluation may be prohibitive, mainly for large sample sizes or period lengths.

In the first step, we have estimated the linear model in Equation 6.14. The coefficient estimates are displayed in Table 6. As expected, we observe negative effects of Saturday and Sunday, which lead to a decrease of $\log(\text{PM}_{10})$ levels during the weekends.

Table 6 – Estimated coefficients of the linear model.

Parameter	μ	α_1	α_2	β_1	β_2
Estimate	3.2650	-0.0921	-0.2579	0.0640	0.0003

The selected orders of the PAR models by the BIC for each estimator are displayed in Table 7. We note that the PAR model with better (smaller) BIC is obtained by the RYWE, which indicates that this estimator provides a good compromise between adjustment and parsimony.

Table 7 – Selected AR orders using the BIC.

Estimator	BIC	p_1	p_2	p_3	p_4	p_5	p_6	p_7
YWE	-2156.25	1	1	1	1	1	1	1
RYWE	-2205.94	1	1	4	2	2	1	1
RLSE	-2200.15	1	1	4	1	2	1	1

The estimates of the AR coefficients provided by YWE, RYWE and RLSE are given in Table 8. It is clear the presence of Periodic Correlation in the data. In general, the order 1 estimates provided by YWE and RLSE agree, while the estimates from RYWE tend to be greater than the other estimators.

Table 8 – Estimates of the AR coefficients for YWE, RYWE and RLSE.

Estimator	i	ν						
		1	2	3	4	5	6	7
$\hat{\phi}_i(\nu)$	1	0.626	0.532	0.485	0.374	0.595	0.312	0.498
$\tilde{\phi}_i(\nu)$	1	0.614	0.482	0.630	0.529	0.595	0.361	0.474
	2	0.000	0.000	-0.107	-0.143	-0.258	0.000	0.000
	3	0.000	0.000	0.451	0.000	0.000	0.000	0.000
	4	0.000	0.000	-0.374	0.000	0.000	0.000	0.000
$\check{\phi}_i(\nu)$	1	0.661	0.513	0.479	0.376	0.638	0.293	0.522
	2	0.000	0.000	-0.018	0.000	-0.167	0.000	0.000
	3	0.000	0.000	0.271	0.000	0.000	0.000	0.000
	4	0.000	0.000	-0.240	0.000	0.000	0.000	0.000

The fitting performance will be accessed through the in-sample Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), symmetric MAPE (sMAPE) and Median of Absolute Deviation (MAD). The RMSE and the MAD are well-known and, for a discussion on MAPE and sMAPE, we refer to Flores (1986). The results are presented in Table 9. We can see that the RLSE provides the better fit in terms of RMSE, this is not surprising since this model has more parameters and the objective function to obtain the estimator is closely related to the Least Square Estimator one. By percentage, the RYWE fit has smaller error (minimum MAPE and sMAPE). Since MAD is resistant to outliers, it indicates that RYWE provides a better fit to standard observations compared to YWE and RLSE.

Table 9 – Fitting performance of the estimated models.

Statistic	Estimator		
	YWE	RYWE	RLSE
RMSE	0.2246	0.2245	0.2235
MAPE	5.3420	5.2545	5.2707
sMAPE	2.6454	2.6024	2.6104
MAD	0.2120	0.1988	0.2078

Figures 9, 10 and 11 present the classic ACF of the residuals of each model. It can be seen that all the models were able to fully explain the correlation structure of the data. Finally, for all models, the residuals have not passed the Jarque-Bera normality test (JARQUE; BERA, 1980), presenting p -values < 0.05 .

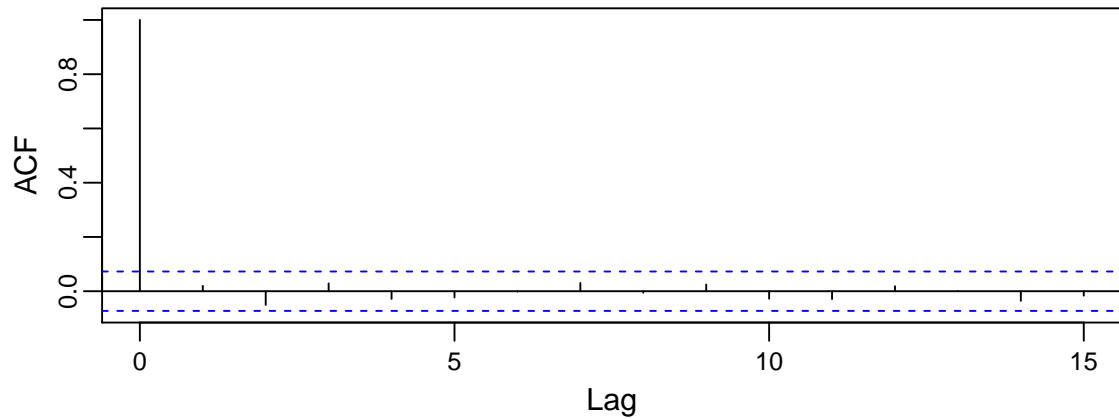


Figure 9 – ACF of the residuals of the YWE fit.

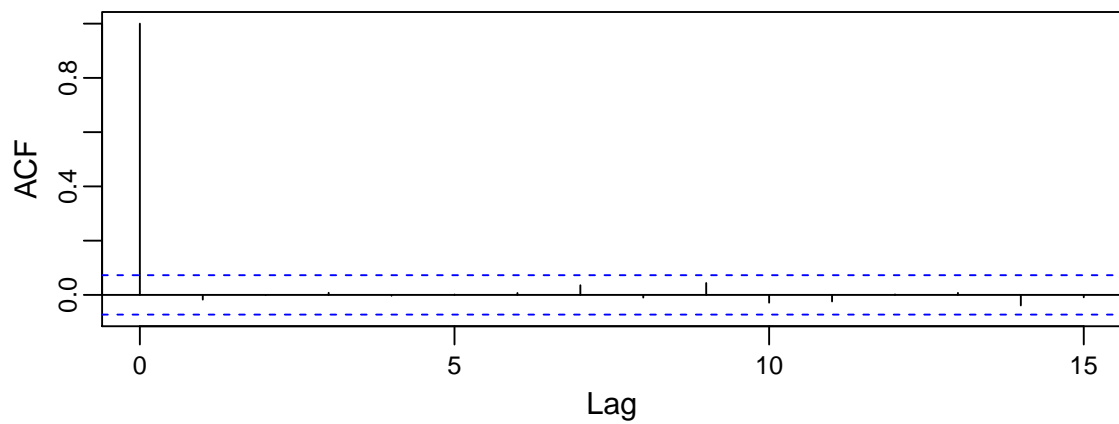


Figure 10 – ACF of the residuals of the RYWE fit.

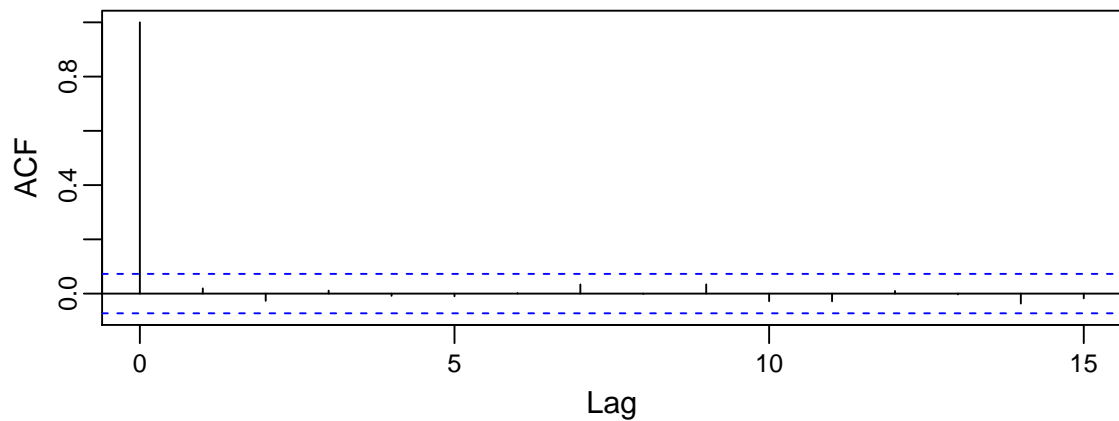


Figure 11 – ACF of the residuals of the RLSE fit.

6.5 Conclusions

This paper reviews different estimation methodologies for PAR models. More specifically, we consider the so-called YWE (MCLEOD, 1994), the RYWE (SARNAGLIA; REISEN; LÉVY-LEDUC, 2010) and the RLSE (SHAO, 2008). The finite sample performance of these methods are compared through a Monte Carlo experiment. The perfor-

mance of RLSE is remarkably good under uncontaminated and contaminated scenarios, even under asymmetric errors, which violates Assumption 3 which ensures its asymptotic normality. The RYWE is quite resistant to outliers, however it has a poor performance under asymmetric errors, mainly under weak correlation scenarios. As expected, YWE empirical distribution is resistant to departures from normality, however this estimator is completely affected by the presence of outliers. In order to illustrate the methodologies considered in this paper, we have analyzed the daily mean PM₁₀ concentrations collected by the monitoring station located at Enseada do Suá, ES, Brazil. The RYWE has provided the more parsimonious model according to BIC. In terms of RMSE, the best fit was provided by the RLSE, while according to MAPE, sMAPE and MAD the best fit was achieved by the RYWE.

6.6 Acknowledgements

The results in this paper are part of the Master thesis of the first author in the PPGEA-UFES under supervision of the second and third authors. The authors gratefully acknowledge partial financial support from Capes/ES, Fapes/ES and CNPq/Brazil.

7 Conclusões e Trabalhos Futuros

Nesta dissertação foram revisadas as metodologias de ajuste de modelos PAR de Yule-Walker clássica (MCLEOD, 1994), de Yule-Walker robusta (SARNAGLIA; REISEN; LÉVY-LEDUC, 2010) e de Mínimos Quadrados Robusta (SHAO, 2008). A robustez em Sarnaglia, Reisen & Lévy-Leduc (2010) é alcançada através da introdução de uma generalização da função de autocovariância robusta (MA; GENTON, 2000) nas equações de Yule-Walker periódicas. Já Shao (2008) faz uso iterativo dos mínimos quadrados em uma versão robusta da série obtida através de uma função ímpar, limitada e diferenciável como a de Huber (HUBER et al., 1964).

Quando comparadas empiricamente através de simulações de Monte Carlo, a performance dos estimadores robustos é muito superior a do estimador clássico em cenário gaussiano contaminado, enquanto que, para cenário gaussiano não contaminado, os estimadores robustos mostraram performances competitivas com as do clássico. Além disso o estimador proposto por Sarnaglia, Reisen & Lévy-Leduc (2010) é muito sensível a desvios de normalidade, já o estimador de Shao (2008) não apresenta essa limitação.

As metodologias comparadas nesta dissertação foram aplicadas na série temporal das médias diárias do poluente MP_{10} coletado pela estação da RAMQAr da Enseada do Suá entre 1 Janeiro de 2014 e 29 de Dezembro de 2015. O estimador mais parcimonioso de acordo com o critério BIC foi o de Yule-Walker robusto. Já em termos de Erro Quadrático Médio, o melhor ajuste foi obtido pelo estimador de Mínimos Quadrados Robusto, enquanto que de acordo com Erro Percentual Absoluto Médio, Erro Percentual Absoluto Médio Simétrico e Desvio Absoluto Médio o melhor ajuste foi atingido pelo estimador de Yule-Walker robusto. Isso indica melhor desempenho das metodologias robustas em comparação com a clássica.

Este trabalho permite ilustrar como perspectivas futuras o estudo da correlação periódica em dados de poluição do ar no domínio da frequência. Em particular, uma futura linha de pesquisa aberta por esta dissertação consistirá da investigação e generalização de métodos de *bootstrap* (PAPARODITIS; POLITIS, 1999) para o contexto de processos periodicamente estacionários. Essas metodologias são amplamente justificadas pelas características periódicas que os dados da RAMQAr apresentam.

Referências Bibliográficas

- ANDERSON, P. L.; MEERSCHAERT, M. M. Parameter estimation for periodically stationary time series. *Journal of Time Series Analysis*, Wiley Online Library, v. 26, n. 4, p. 489–518, 2005. 34
- BAIRD, C. *Química ambiental*. [S.l.]: Reverté, 2001. 13
- BASAWA, I.; LUND, R. Large sample properties of parameter estimates for periodic arma models. *Journal of Time Series Analysis*, Wiley Online Library, v. 22, n. 6, p. 651–663, 2001. 20, 34, 37, 39
- BLOOMFIELD, P.; HURD, H. L.; LUND, R. B. Periodic correlation in stratospheric ozone data. *Journal of Time Series Analysis*, Wiley Online Library, v. 15, n. 2, p. 127–150, 1994. 16, 18, 19, 33, 35
- BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 211–252, 1964. 28
- BOX, G. E.; JENKINS, G. M. *Time series analysis: forecasting and control, revised ed.* [S.l.]: Holden-Day, 1976. 24
- BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden day, 1970. 14, 18, 21, 34
- BRAGA, B. et al. *Introdução à Engenharia Ambiental: o Desafio do Desenvolvimento Sustentável, 2ª edição*. [S.l.]: São Paulo: Editora Pearson, 2005. 13
- BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods*. 2nd. ed. New York, NY: Springer Science, 2006. (Springer Series in Statistics). 18, 21, 33
- BROSZKIEWICZ-SUWAJ, E.; MAKAGON, A.; WERON, R.; WYŁOMĄŃSKA, A. On detecting and modeling periodic correlation in financial data. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 336, n. 1, p. 196–205, 2004. 19
- FLORES, B. E. A pragmatic view of accuracy measurement in forecasting. *Omega*, Elsevier, v. 14, n. 2, p. 93–98, 1986. 50
- FOX, A. J. Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 350–363, 1972. 15, 20
- FRANSES, P. H.; PAAP, R. *Forecasting with periodic autoregressive time-series models*. [S.l.]: Oxford, UK, Blackwell Publishers Ltd, 2002. 18, 19
- FRANSES, P. H.; PAAP, R. *Periodic time series models*. [S.l.]: OUP Oxford, 2004. 18
- GARDNER, W. A.; FRANKS, L. E. Characterization of cyclostationary random signal processes. *Information Theory, IEEE Transactions on*, IEEE, v. 21, n. 1, p. 4–14, 1975. 18, 33
- GARDNER, W. A.; NAPOLITANO, A.; PAURA, L. Cyclostationarity: Half a century of research. *Signal processing*, Elsevier, v. 86, n. 4, p. 639–697, 2006. 18

- GLADYSHEV, E. G. Periodically correlated random sequences. *Sov. Math.*, v. 2, p. 385–388, 1961. 14, 18, 33
- GODISH, T. *Air quality*. [S.l.]: Boca Raton: CRC Press, LLC, 1997. 13
- HOLGATE, S. T.; KOREN, H. S.; SAMET, J. M.; MAYNARD, R. L. *Air pollution and health*. [S.l.]: Academic Press, 1999. 13
- HUBER, P. J. et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 35, n. 1, p. 73–101, 1964. 20, 53
- HURD, H. L.; GERR, N. L. Graphical methods for determining the presence of periodic correlation. *Journal of Time Series Analysis*, Wiley Online Library, v. 12, n. 4, p. 337–350, 1991. 19, 33, 35
- HURD, H. L.; MIAMEE, A. *Periodically correlated random sequences: Spectral theory and practice*. [S.l.]: John Wiley & Sons, 2007. v. 355. 19
- IBGE. *Sinopse do Censo Demográfico 2010*. Rio de Janeiro, RJ, 2011. 30
- IEMA. *Relatório da Qualidade do Ar da Região da Grande Vitória - 2005*. Espírito Santo, ES, 2013. 30
- IJSN. *Perfil Regional - Região Metropolitana da Grande Vitória*. Espírito Santo, ES, 2008. 30
- JACOBSON, M. *Atmospheric Pollution: History, Science, and Regulation*. Cambridge University Press, 2002. ISBN 9780521010443. Disponível em: <https://books.google.com.br/books?id=NN5S0\3dEvkC>. 13
- JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, Elsevier, v. 6, n. 3, p. 255–259, 1980. 50
- JONES, R. H.; BRELSFORD, W. M. Time series with periodic structure. *Biometrika*, Biometrika Trust, v. 54, n. 3-4, p. 403–408, 1967. 19
- LUND, R.; BASAWA, I. Asymptotics, nonparametrics, and time series. In: _____. [S.l.]: CRC Press, 1999. cap. Modeling and Inference for Periodically Correlated Time Series, p. 37. 38
- LUND, R.; BASAWA, I. Recursive prediction and likelihood evaluation for periodic arma models. *Journal of Time Series Analysis*, Wiley Online Library, v. 21, n. 1, p. 75–93, 2000. 20
- LUND, R. B.; HURD, H. L.; BLOOMFIELD, P.; SMITH, R. L. Climatological time series with periodic correlation. v. 8, p. 2787–2809, 1995. 33
- LÜTKEPOHL, H.; POSKITT, D. S. Estimating orthogonal impulse responses via vector autoregressive models. *Econometric Theory*, Cambridge Univ Press, v. 7, n. 04, p. 487–496, 1991. 19

- MA, Y.; GENTON, M. G. Highly robust estimation of the autocovariance function. *Journal of time series analysis*, Wiley Online Library, v. 21, n. 6, p. 663–684, 2000. 20, 33, 38, 53
- MARONNA, R.; MARTIN, R. D.; YOHAI, V. *Robust statistics*. [S.l.]: John Wiley & Sons, Chichester. ISBN, 2006. 20
- MARTIN, D. E.; KEDEM, B. Estimation of the period of periodically correlated sequences. *Journal of Time Series Analysis*, Wiley Online Library, v. 14, n. 2, p. 193–205, 1993. 19
- MCLEOD, A. Diagnostic checking of periodic autoregression models with application. *Journal of Time Series Analysis*, v. 15, n. 2, p. 221–223, 1994. 8, 9, 16, 18, 19, 20, 33, 48, 51, 53
- NOAKES, D. J.; MCLEOD, A. I.; HIPEL, K. W. Forecasting monthly riverflow time series. *International Journal of Forecasting*, Elsevier, v. 1, n. 2, p. 179–190, 1985. 34
- PAPARODITIS, E.; POLITIS, D. N. The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, Wiley Online Library, v. 20, n. 2, p. 193–222, 1999. 53
- PRIESTLEY, M. B. *Spectral Analysis and Time Series*. [S.l.]: Academic press, 1981. 18, 21, 33
- REISEN, V. A.; SARNAGLIA, A. J. Q.; JR, N. C. Reis; LÉVY-LEDUC, C.; SANTOS, J. M. Modeling and forecasting daily average PM₁₀ concentrations by a seasonal long-memory model with volatility. *Environmental Modelling & Software*, v. 51, p. 286–95, 2014. 34
- REISEN, V. A.; SILVA, A. N. *O uso da linguagem R para cálculos de estatística básica*. [S.l.]: EDUFES, 2011. 32
- ROUSSEEUW, P. J.; CROUX, C. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, Taylor & Francis Group, v. 88, n. 424, p. 1273–1283, 1993. 20, 38, 49
- SAKAI, H. Circular lattice filtering using pagano’s method. *IEEE transactions on acoustics, speech, and signal processing*, v. 30, n. 2, p. 279–287, 1982. 20, 35
- SARNAGLIA, A. J. Q.; REISEN, V. A.; BONDON, P. Periodic ARMA models: Application to particulate matter concentrations. In: IEEE. *Signal Processing Conference (EUSIPCO), 2015 23rd European*. [S.l.], 2015. p. 2181–2185. 16, 34
- SARNAGLIA, A. J. Q.; REISEN, V. A.; BONDOU, P.; LÉVY-LEDUC, C. A robust estimation approach for fitting a PARMA model to real data. In: IEEE. *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. [S.l.], 2016. p. 1–5. 16, 34
- SARNAGLIA, A. J. Q.; REISEN, V. A.; LÉVY-LEDUC, C. Robust estimation of periodic autoregressive processes in the presence of additive outliers. *Journal of multivariate analysis*, Elsevier, v. 101, n. 9, p. 2168–2183, 2010. 8, 9, 15, 16, 20, 33, 34, 36, 37, 38, 39, 51, 53
- SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. 48

- SHAO, Q. Robust estimation for periodic autoregressive time series. *Journal of Time Series Analysis*, Wiley Online Library, v. 29, n. 2, p. 251–263, 2008. 8, 9, 16, 20, 33, 34, 38, 39, 40, 41, 51, 53
- SHAO, Q.; LUND, R. Computation and characterization of autocorrelations and partial autocorrelations in periodic arma models. *Journal of Time Series Analysis*, Wiley Online Library, v. 25, n. 3, p. 359–372, 2004. 35
- SOUZA, J. B. et al. Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data. *In Press, JRSS-Series C*, JRSS, 2017. 34
- SOUZA, J. B. d.; REISEN, V. A.; SANTOS, J. M.; FRANCO, G. C. Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. *Revista de saude publica*, SciELO Public Health, v. 48, n. 3, p. 451–458, 2014. 14
- TIAO, G.; GRUPE, M. Hidden periodic autoregressive-moving average models in time series data. *Biometrika*, JSTOR, p. 365–373, 1980. 14, 29, 33
- TROUTMAN, B. M. Some results in periodic autoregression. *Biometrika*, Biometrika Trust, v. 66, n. 2, p. 219–228, 1979. 19
- VECCHIA, A.; BALLERINI, R. Testing for periodic autocorrelations in seasonal time series data. *Biometrika*, Biometrika Trust, v. 78, n. 1, p. 53–63, 1991. 18, 19
- WANG, J.; CHEN, T.; HUANG, B. Cyclo-period estimation for discrete-time cyclo-stationary signals. *IEEE transactions on signal processing*, IEEE, v. 54, n. 1, p. 83–94, 2006. 19
- WEI, W. W. et al. *Time series analysis: univariate and multivariate methods*. [S.l.]: Pearson Addison Wesley, 2006. 18, 21
- WHO - World Health Organization. *WHO Air quality guidelines for particulate matter and ozone and nitrogen dioxide and sulfur dioxide*. [S.l.], 2005. 13, 14
- WOLD, H. *A study in the analysis of stationary time series*. Tese (Doutorado) — Almqvist & Wiksell, 1938. 24